ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

Evaluating Social Programs

June 15 – June 19, 2015

Executive Education at the Jameel Poverty Action Lab

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

Table of Contents

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

PROGRAM

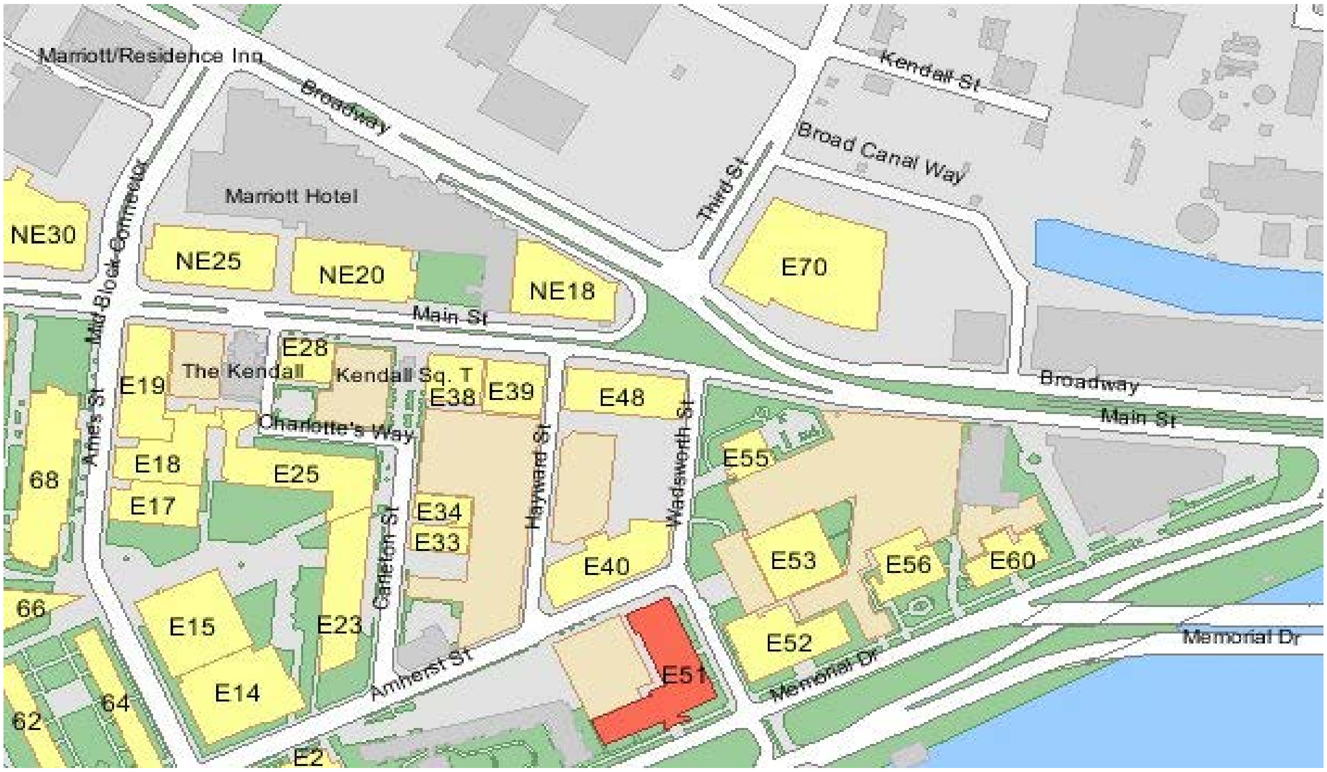
J-PAL Executive Education Course in Evaluating Social Programs, June 15 – 19, 2015

Massachusetts Institute of Technology

	Monday **June 15, 2015**	**Tuesday** **June 16, 2015**	**Wednesday** **June 17, 2015**	**Thursday** **June 18, 2015**	**Friday** **June 19, 2015**
8:30 – 9:00	Registration/Breakfast	Breakfast	Breakfast	Breakfast	Breakfast
9:00 – 10:30	Welcoming remarks and Expectations Survey Lecture 1*: What is Evaluation* *(Marc Shotland)*	Lecture 3: *Why Randomize* *(Dan Levy)*	Lecture 5: *Sampling and Sample Size* *(Ben Olken)*	Lecture 6: *Threats and Analysis* *(Shawn Cole)*	Lecture 8: *Project from Start to Finish* *(Rachel Glennerster)*
10:30 – 10:45	Coffee Break	Coffee Break	Coffee Break	Coffee Break	Coffee Break
10:45 – 12:00	Group work on case study 1: Theory of Change: *Women as Policymakers* Decision on group project	Group Exercise A: *Random Sampling* Group work on presentation: Indicators	Group Exercise C*:* *Sample Size Estimation*	Group work on presentation: Threats and Analysis	Feedback survey Group presentations
12:00 – 1:00	Lunch	Lunch	Lunch	Lunch	Lunch
1:00 – 2:30	Lecture 2: *Outcomes, Impact, and Indicators* *(Shawn Powers)*	Lecture 4: *How to Randomize* *(Joe Doyle)*	Group work on case study 4: Threats and Analysis: *Deworming in Kenya*	*Lecture 7: Generalizability* *(Rachel Glennerster)*	Group presentations Closing remarks
2:30 – 3:00	Coffee Break	Coffee Break	Coffee Break	Coffee Break	
3:00 – 4:00	Group work on presentation: Theory of change, research question	Group Exercise B: *Randomization Mechanics*	Group work on presentation: Randomization Design	Group work on presentation: Finalize presentation	
4:00 – 5:00	Group work on case study 2: Why Randomize: *Learn to Read*	Group work on case study 3: How to Randomize: *Extra Teacher Program*	Group work on presentation: Power and sample size		
7:00-9:00	Dinner: EVOO				

Course Location

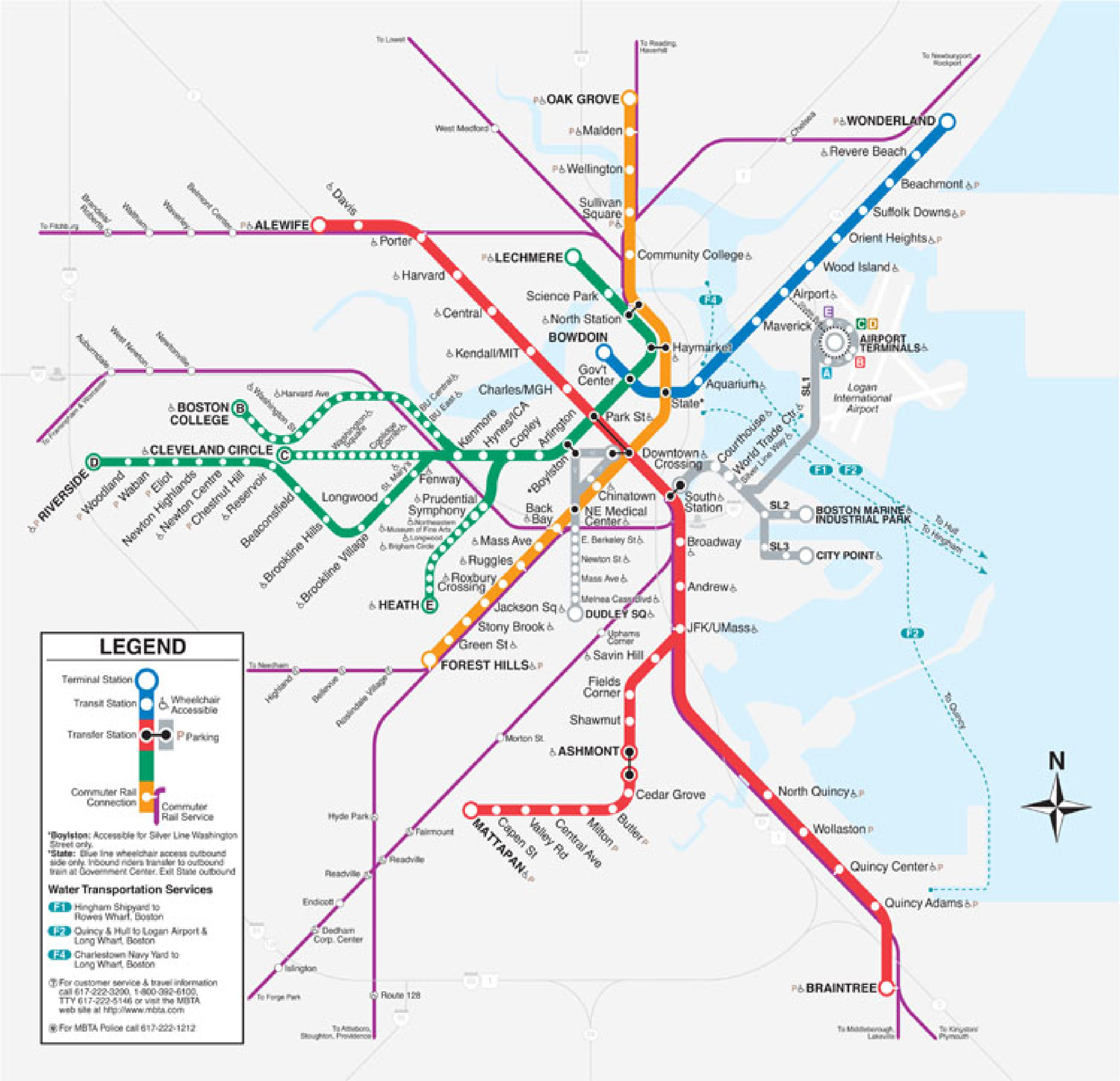
The executive education course will be held in Building E51 (the Tang Center) on the east side of MIT's campus (in the Kendall Square area) in Cambridge, MA. The Tang Center is located at the corner of Wadsworth Street and Amherst Street. Lectures will be held in room E51-395 on the third floor of the building.



Directions from the Boston/Logan airport to the Kendall Square area:

By subway: Direct bus service from the airport to the subway system is located on the ground transportation level of the airport. Follow signs to *Silver Line* bus service, which takes you to *South Station* located on the red line. Take the *outbound* red line train to the *Kendall/MIT* station. Total one-way cost: $2.00

Alternatively, free shuttle bus service from the airport to the subway system is available on the ground transportation level of the airport. Follow signs for *Massport Shuttle* service, which takes you to the *Airport* station located on the blue line. Take the *inbound* blue line train to *Government Center* and switch to a green line train (B, C, D, or E). Take the green line train to the next stop, which is *Park Street* and switch to the red line, going *outbound* towards *Alewife*. Take the *outbound* red line train to the *Kendall/MIT* station. Total one-way cost: $2.00

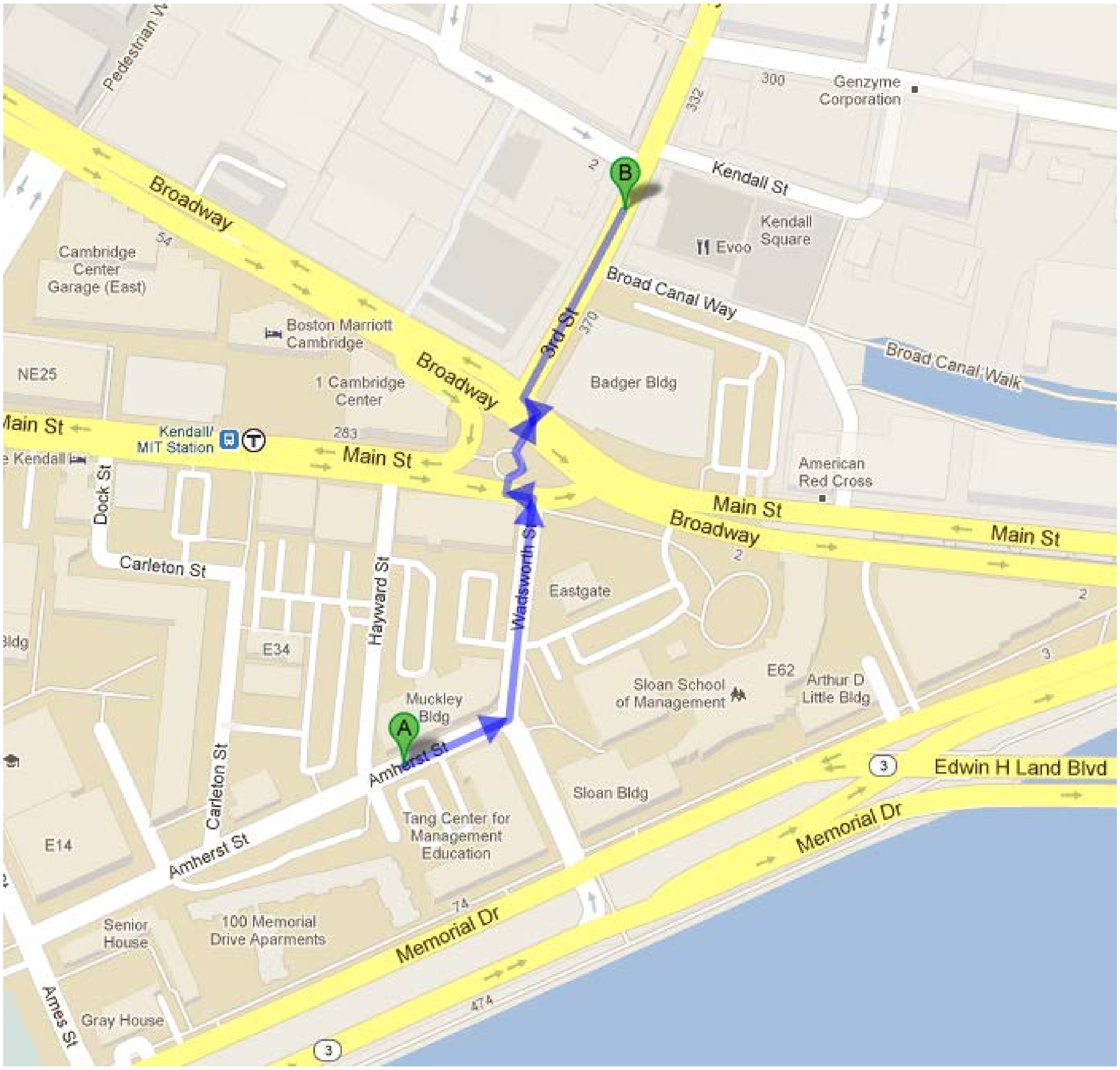


By cab: Ask your driver to take the Storrow Drive route towards Kendall Square.
(Cab fare approximately $30.00)

Driving directions: Take Callahan Tunnel to 93 North to Exit 26 Cambridge/Storrow Drive. Follow the signs to Storrow Drive and take the Kendall Square/ Government Center exit (on left). At the end of the ramp, bear right towards Kendall Square.

Directions to dinner at EVOO:

Walking directions: 350 3rd St, Cambridge, MA 02142



Course Objectives

Our executive training program is designed for people from a variety of backgrounds: managers and researchers from international development organisations, foundations, governments and non-governmental organisations from around the world, as well as trained economists looking to retool.

The course is a **full-time course**. It is important for participants to **attend all lectures and group work** in order to successfully complete the course and receive the certificate of completion.

Course Coverage

The following key questions and concepts will be covered:

- Why and when is a rigorous evaluation of social impact needed?
- The common pitfalls of evaluations, and how randomization can help.
- The key components of a good randomized evaluation design
- Alternative techniques for incorporating randomization into project design.
- How do you determine the appropriate sample size, measure outcomes, and manage data?
- Guarding against threats that may undermine the integrity of the results.
- Techniques for the analysis and interpretation of results.
- How to maximise policy impact and test external validity.

The program will achieve these goals through a diverse set of integrated teaching methods. Expert researchers will provide both theoretical and example-based classes complemented by workshops where participants can apply key concepts to real world examples.

By examining both successful and problematic evaluations, participants will better understand the significance of specific details of randomized evaluations. Furthermore, the program will offer extensive opportunities to apply these ideas ensuring that participants will leave with the knowledge, experience, and confidence necessary to conduct their own randomized evaluations.

J-PAL Lecturers

Shawn Cole
Associate Professor of Business Administration
Harvard Business School

Shawn Cole is an Associate Professor in the Finance Unit at Harvard Business School. His research examines corporate finance, banking, and consumer finance in developing countries, covering topics such as bank competition, government regulation, and household investment decisions. He has conducted randomized evaluations in education and financial literacy, as well as evaluations of market-based products to help farmers manage risk.

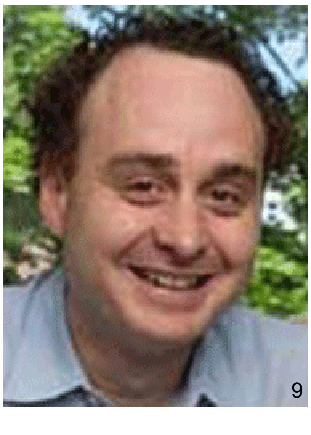
Joseph Doyle
Associate Professor of Management
Massachusetts Institute of Technology

Joseph Doyle is the Erwin H. Schell Professor of Management and a Professor of Applied Economics at the MIT Sloan School of Management. His research focuses in particular on child welfare and health. He studies the effects of foster care and other interventions on child outcomes, the returns on healthcare spending, and the role of health insurance on the quality of treatment provided to patients and health outcomes. Doyle received a Ph.D. in Economics from the University of Chicago in 2002.

Rachel Glennerster
Executive Director
J-PAL Global

Rachel Glennerster is Executive Director of J-PAL. Her current research includes randomized evaluations of community driven development and adoption of new agricultural technologies in Sierra Leone, empowerment of adolescent girls in Bangladesh, and health, education, and microfinance in India. She oversees J-PAL’s work to translate research findings into policy action and helped establish Deworm the World.

Dan Levy
Lecturer in Public Policy
Harvard Kennedy School of Government



Dan Levy is a Lecturer in Public Policy at Harvard University's John F. Kennedy School of Government. He is currently involved in evaluations of a conditional cash transfer program in Jamaica and a set of education interventions in Burkina Faso. He also provides technical assistance and training to Mexico's Social Development Ministry on impact evaluations of social programs.

Benjamin Olken
Professor of Economics
Massachusetts Institute of Technology

Benjamin Olken is a Professor in the Department of Economics at MIT. In addition, he is a Director of J-PAL and a member of the Executive Committee. His research focuses on the political economy of developing countries with a particular focus on corruption. He is involved in several randomized evaluations in Indonesia that seek to reduce corruption and improve targeting of programs that provide local public goods to villages..

Shawn Powers
Senior Policy Manager
J-PAL Gobal

Shawn Powers joined J-PAL in 2011 and manages J-PAL's Education program, which includes the Post-Primary Education Initiative. He is the author or co-author of a number of J-PAL publications, including review papers on primary and post-primary education. Shawn speaks regularly to representatives of international organizations, foundations, governments, and NGOs to help them understand and apply rigorous evidence in education and other sectors. He also facilitates partnerships between researchers and implementing organizations.

Marc Shotland
Director of Training and Senior Research Manager
J-PAL Global

Marc holds an MPA/International Development degree from Harvard University's Kennedy School of Government and a Bachelors degree in Economics from Williams College. He first joined Professors Duflo and Banerjee in the summer of 2002 to run randomized evaluations of education interventions as a field research associate in India. In 2004 he joined the Poverty Action Lab's Cambridge office as a research manager. He left in 2006 to earn his Masters at Harvard before rejoining J-PAL in 2008 in his current position.

List of Participants

Sr. No.	Name	Organization Name	Country
1	Abidemi Coker	University of Jyvaskyla	Finland
2	Aina Rundgren	Orebro Municipality	Sweden
3	Alan Rico	Escuela Superior de Economia y Negocios (ESEN)	El Salvador
4	Amitpal Singh	New York University	United States
5	Bohwa Lee	Korea Institute of Public Finance	Korea
6	Christiane Lorenz	Ernst Basler + Partner	Switzerland
7	Dawit Mekonnen	International Food Policy Research Institute	United States
8	Deniz Sanin	Duke University	United States
9	Eman Tarawneh	We Love Reading	Jordan
10	Emrul Hasan	Plan International Canada	Canada
11	Jason Bauman	J-PAL	United States
12	Jeneen Garcia	Global Environment Facility Independent Evaluation Office	United States
13	Koba Turmanidze	Caucasus Research Resource Center (CRRC)	Georgia
14	Malin Bengtsson	Municipal of Norrkoping	Sweden
15	Maria Fernanda P S Mendes	Credit Suisse Hedging Griffo Institute	Brazil
16	Michelle Woodford	J-PAL	United States
17	Mihye Heo	Center for Performance Evaluation & Management	Korea
18	Naoko Uchiyama	Research Institute for Economics & Business Administration, Kobe University	Japan
19	Nicholas Sabin	University of Oxford (Said Business School)	United Kingdom
20	Nicole Pollock	City of Providence	United States
21	Paul Scott	ASME	United States
22	Sofie Sjoborg	The Swedish Association Of Local Authorities And Regions (Salar)	Sweden
23	Steve Ryan	J-PAL	United States
24	Suh Yoon Kang	UNICEF India	India
25	Suhaib Al-Absi	We Love Reading	Jordan

26	Thomas Abt	Harvard Kennedy School (HKS)	United States
27	Tomas Bokstrom	Swedish Association of Local Authorities And Regions	Sweden
28	Youngmin Oh	Korea Institute of Public Finance	Korea
29	Zurab Simonia	Millennium Challenge Account, Georgia	Georgia

Groups

Group 1
TA: Alison Fahey
Room: 390

Alan Rico
Bohwa Lee
Jeneen Garcia
Maria Fernanda P S Mendes
Mihye Heo

Group 2
TA: Arianna Ornaghi
Room: 376

Deniz Sanin
Koba Turmanidze
Naoko Uchiyama
Suhaib Al-Absi
Zurab Simonia

Group 3
TA: Geetika Mehra
Room: 372

Christiane Lorenz
Emrul Hasan
Jason Bauman
Tomas Bokstrom
Youngmin Oh

Group 4
TA: Ariella Park
Room: 151

Amitpal Singh
Eman Tarawneh
Sofie Sjoborg
Steve Ryan

Group 5
TA: Julia Chabrier
Room: 149

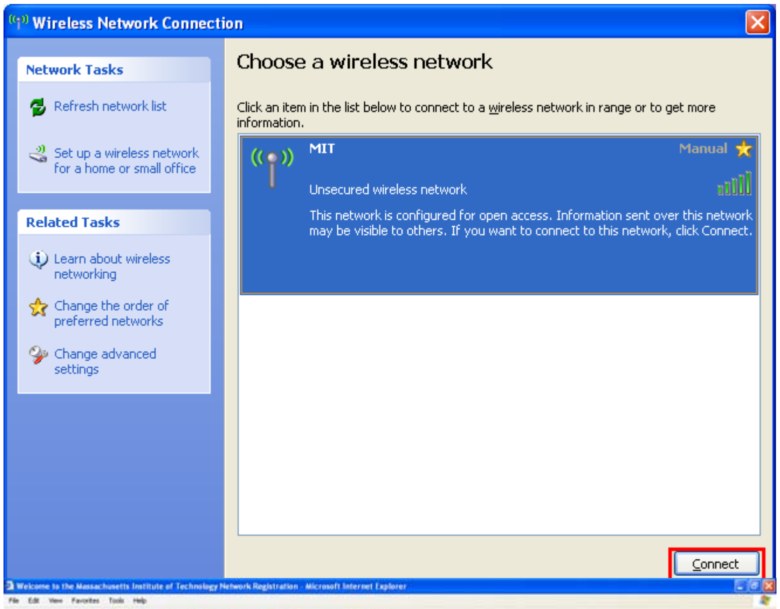
Aina Rundgren
Malin Bengtsson
Nicole Pollock
Paul Scott
Thomas Abt

Group 6
TA: Justin de Benedictis-Kessner
Room: 063

Abidemi Coker
Dawit Mekonnen
Michelle Woodford
Nicholas Sabin
Suh Yoon Kang

MIT Wireless Instructions

1) Search for available wireless networks. Select "MIT" and click connect. There are no passwords required here.



Wireless Network Connection

Choose a wireless network

Click an item in the list below to connect to a wireless network in range or to get more information.

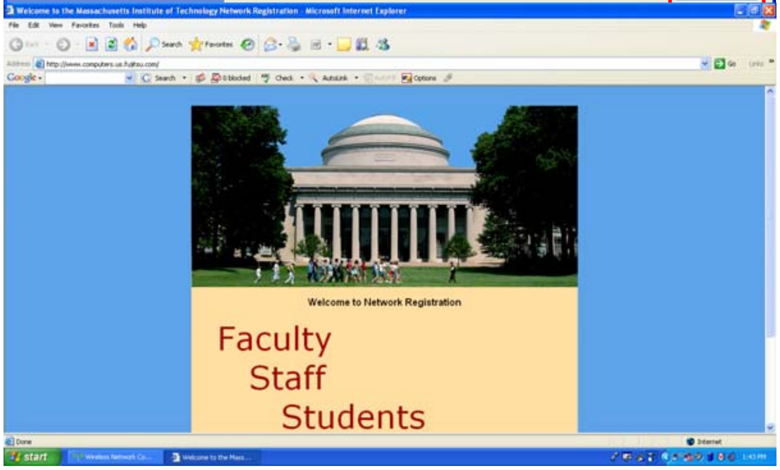
MIT

Unsecured wireless network

This network is configured for open access. Information sent over this network may be visible to others. If you want to connect to this network, click Connect.

Network Tasks
Refresh network list
Set up a wireless network for a home or small office
Related Tasks
Learn about wireless networking
Change the order of preferred networks
Change advanced settings
Manual
Connect

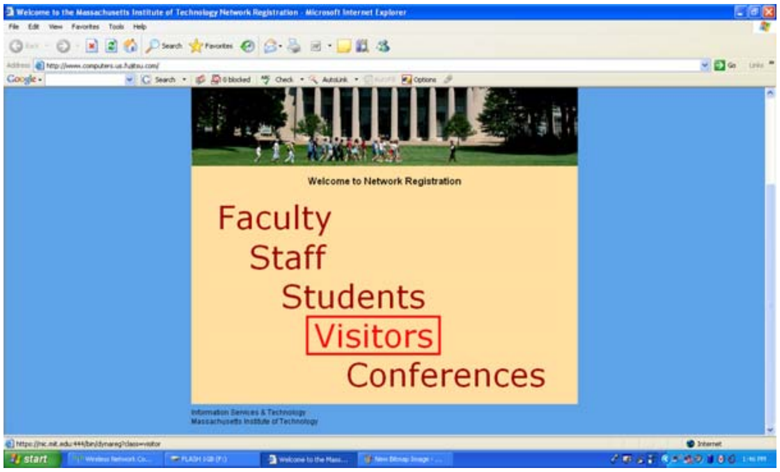
2) You will be automatically redirected to a screen that looks like this when you open a web browser.



Welcome to the Massachusetts Institute of Technology Network Registration - Microsoft Internet Explorer

Welcome to Network Registration

3) Select the visitor's option.



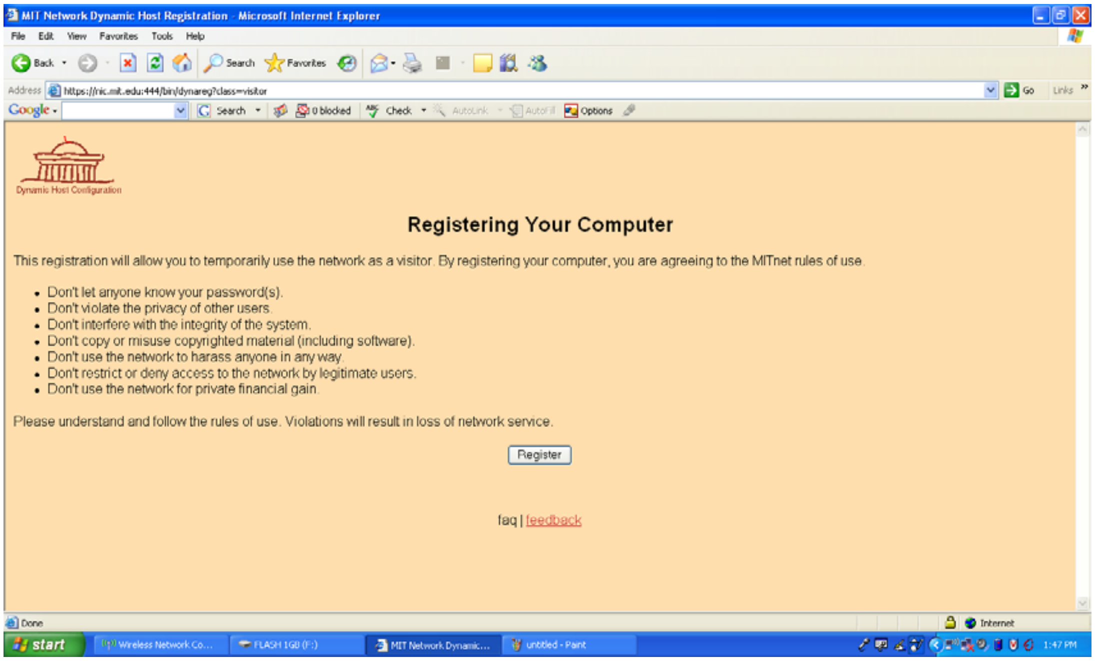
Welcome to the Massachusetts Institute of Technology Network Registration - Microsoft Internet Explorer

Welcome to Network Registration

Faculty
Staff
Students
Visitors
Conferences
Information Services & Technology
Massachusetts Institute of Technology
start

MIT Wireless Instructions

4) After reviewing the guidelines click the register button.


MIT Network Dynamic Host Registration - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address https://nic.mit.edu:444/bin/dynareg?class=visitor
Go Links
Google Search 0 blocked Check AutoLink AutoFill Options
Dynamic Host Configuration
Done
Internet
start
Wireless Network Co...
FLASH 1GB (F:)
MIT Network Dynamic...
untitled - Paint
1:47 PM

Registering Your Computer

This registration will allow you to temporarily use the network as a visitor. By registering your computer, you are agreeing to the MITnet rules of use.

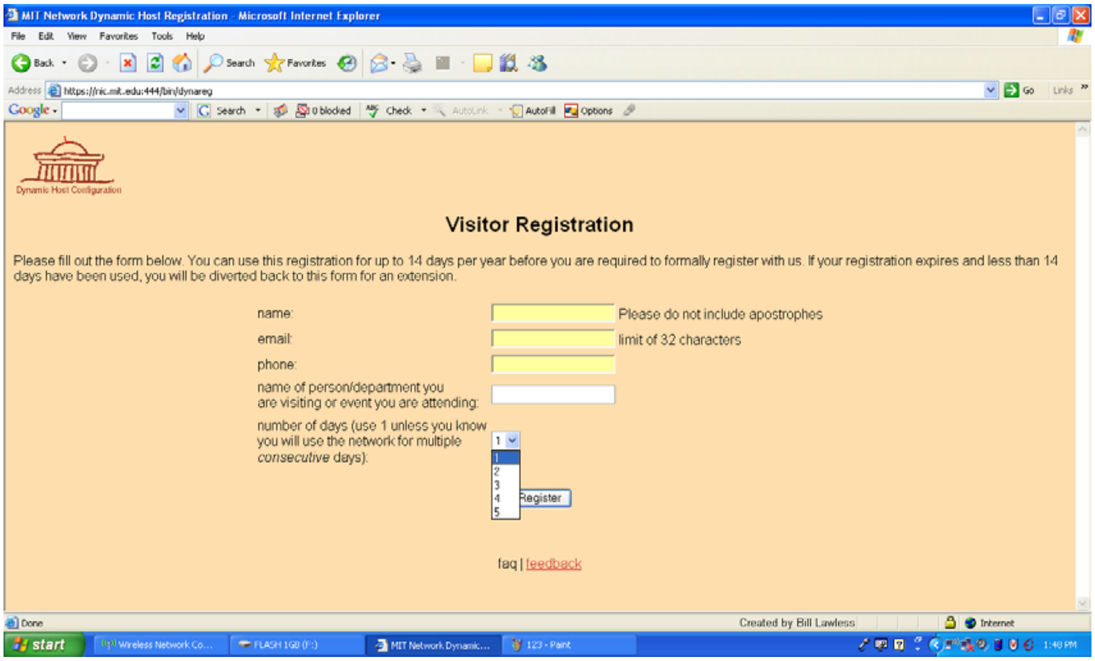
- Don't let anyone know your password(s).
- Don't violate the privacy of other users.
- Don't interfere with the integrity of the system.
- Don't copy or misuse copyrighted material (including software).
- Don't use the network to harass anyone in any way.
- Don't restrict or deny access to the network by legitimate users.
- Don't use the network for private financial gain.

Please understand and follow the rules of use. Violations will result in loss of network service.

Register

faq | feedback

5. Fill out the form. Select the number of days that you will be here (5). Click the register button to submit.


MIT Network Dynamic Host Registration - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address https://nic.mit.edu:444/bin/dynareg
Go Links
Google Search 0 blocked Check AutoLink AutoFill Options
Dynamic Host Configuration
Done
Created by Bill Lawless
Internet
start
Wireless Network Co...
FLASH 1GB (F:)
MIT Network Dynamic...
123 - Paint
1:49 PM

Visitor Registration

Please fill out the form below. You can use this registration for up to 14 days per year before you are required to formally register with us. If your registration expires and less than 14 days have been used, you will be diverted back to this form for an extension.

name: Please do not include apostrophes

email: limit of 32 characters

phone:

name of person/department you are visiting or event you are attending:

number of days (use 1 unless you know you will use the network for multiple *consecutive* days): 1 (1, 2, 3, 4, 5)

Register

faq | feedback

6) Allow 15 minutes for information to replicate, and you should be all ready to surf the World Wide Web.

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 1: Women as Policymakers

Measuring the effects of political reservations

Thinking about measurement and outcomes

This case study is based on: Raghabendra Chattopadhyay and Esther Duflo, 2004a, "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica* 72(5) 1409-1443.

J-PAL thanks the authors for allowing us to use their paper.

Key vocabulary

Hypothesis: a proposed explanation of and for the effects of a given intervention. Hypotheses are intended to be made *ex ante* or prior to the implementation of the intervention.

Indicators: metrics used to quantify and measure specific short-term and long-term effects of a program

Logical Framework: a management tool used to facilitate the design, execution, and evaluation of an intervention. It involves identifying strategic elements (inputs, outputs, outcomes, and impact) and their causal relationships, indicators, and the assumptions and risks that may influence success or failure.

Theory of Change: describes a strategy or blueprint for achieving a given long-term goal. It identifies the preconditions, pathways, and interventions necessary for an initiative's success.

Introduction

India amended its federal constitution in 1992, devolving power over local development programs from the states to rural councils, or *gram panchayats* (village councils). The village councils now choose which development programs to undertake and how much of the budget to invest in them. The states are also required to reserve a third of village council seats and chairperson positions for women. In most states, the schedule on which different villages must reserve seats and positions is determined randomly. This creates the opportunity to rigorously assess the impact of quotas on politics and government: Do the policies differ when there are more women in government? Do the policies chosen by women in power reflect the policy priorities of women? Since randomization was part of the Indian government program itself, the evaluation planning centered on collecting the data needed to measure impact. The researchers then considered what data to collect and which data collection instruments to use.

Empowering the *panchayati raj*

Village councils, known locally as *panchayat*s, have a long tradition in rural India. Originally, *panchayats* were assemblies *(yat)* of five *(panch)* elders, chosen by the community, convened to mediate disputes between people or villages. In modern times village councils have been formalized into institutions of local self-government.

This formalization came about through the constitution. In 1992, India enacted the 73rd amendment, which directed the states to establish a three-tier *panchayati raj* system. The village council is the grassroot unit[1] of this system, with each council consisting of councilors elected every five years. The councilors elect from among themselves a chairperson called a *pradhan*. Decisions are made by a majority vote and the chairperson has no veto power. But as the only councilor with a full-time appointment, the chairperson wields effective power.

The 73rd amendment aimed to decentralize the delivery of public goods and services essential for development in rural areas. The states were directed to delegate the power to plan and implement local development programs to the village councils. Funds still come from the central government but are no longer earmarked for specific uses. Instead, the village council decides which programs to implement and how much to invest in them. As of 2005, Village Councils can chose programs from 29 specified areas, including welfare services (e.g., public assistance for widows, care for the elderly, maternity care, antenatal care, and child health) and public works (e.g., drinking water, roads, housing, community buildings, electricity, irrigation, and education).

Empowering women in the *panchayati raj*

The village councils are large and diverse. In West Bengal, for example, each council represents up to 12 villages and up to 10,000 people who may vary by religion, ethnicity, caste, and, of course, gender.

Political voice varies by group identities drawn along these lines. If policy preferences vary by group identity and if the councilors' identities influence policy choices, then groups underrepresented in politics and government could be shut out as village councils could ignore those groups' policy priorities. There were fears that the newly empowered village councils would undermine the development priorities of traditionally marginalized groups such as women. To remedy this, the 73rd amendment included two mandates to ensure that investments reflected the needs of everyone in the village council.

The first mandate secures community input. If village council investments are to reflect a community's priorities, the councilors must first know what those priorities are. Accordingly, village councils are required to hold a general assembly every six months or every year to report on activities in the preceding period and to submit the proposed budget to the community for ratification. In addition, the chairpersons are required to set up regular office hours to allow constituents to formally request services and lodge complaints. Both requirements allow constituents to articulate their policy preferences.

The second mandate secures representation in the council for women. States are required to reserve at least a third of all council seats and chairperson positions for women. Furthermore, states must ensure that the seats reserved for women are "allotted by rotation to different constituencies in a *panchayat* (village council)" and that the chairperson positions reserved for women are "allotted by rotation to different *panchayat*s." In other words, they have to

[1] Village councils, called *gram panchayats*, form the basic units of the *panchayat raj*. Village council chairs, elected by the members of the village council, serve as members of the block–subdistrict council (*panchayat samiti*). At the top of the system is the district council (*zilla parishad*) made up of the block–council chairs.

ensure that reserved seats and chairperson positions rotate evenly within and across the village councils.

Randomized quotas in India: What can they teach us?

Your evaluation team has been entrusted with the responsibility to estimate the impact of quotas for women in the village councils. Your evaluation should address all dimensions in which quotas for women are changing local communities in India. What could these dimensions be? What data will you collect? What instruments will you use?

As a first step you want to understand all you can about the quota policy. What needs did it address? What are the pros and cons of the policy? What can we learn from it?

DISCUSSION TOPIC 1

Gender quotas in the village councils

1. What were the main goals of the village councils?
2. Women are underrepresented in politics and government. Only 10 percent of India's national assembly members are women, compared to 17 percent worldwide. Does it matter that women are underrepresented? Why and why not?
3. What were the framers of the 73rd amendment trying to achieve when they introduced quotas for women?

Gender quotas have usually been followed by dramatic increases in the political representation of women. Rwanda, for example, jumped from 24th place in the "women in parliament" rankings to first place (49 percent) after the introduction of quotas in 1996. Similar changes have been seen in Argentina, Burundi, Costa Rica, Iraq, Mozambique, and South Africa. Indeed, as of 2005, 17 of the top 20 countries in the rankings have quotas.

Imagine that your group is the national parliament of a country deciding whether to adopt quotas for women in the national parliament. Randomly divide your group into two parties, one against and one for quotas.

What data to collect

First, you need to be very clear about the likely impact of the program. It is on those dimensions that you believe will be affected that you will try to collect data. What are the main areas in which the quota policy should be evaluated? In which areas do you expect to see a difference as a result of quotas?

What are all the possible effects of quotas?

DISCUSSION TOPIC 2

Using a logical framework to delineate your intermediate and final outcomes of interest

1. Brainstorm the possible effects of quotas, positive, negative, and no effects.
2. What evidence would you collect to strengthen the case of those who are for or against quotas? For each potential effect on your list, also list the indicator(s) you would use for that effect. For example, if you say that quotas will affect political participation of women, the indicator could be "number of women attending the General Assembly."

Multiple outcomes are difficult to interpret, so define a hypothesis

Quotas for women could produce a large number of outcomes in different directions. For example, they may improve the supply of drinking water and worsen the supply of irrigation. Without an *ex ante* hypothesis on the direction in which these different variables should be affected by the quota policy, it will be very difficult to make sense of any result we find. Think of the following: if you take 500 villages and randomly assign them in your computer to a "treatment" group and a "control" group, and then run regressions to see whether the villages look different along a hundred outcomes, would you expect to see some differences among them? Would it make sense to rationalize those results *ex post?*

The same applies to this case: if you present your report in front of the commission that mandated that you evaluate this policy, explaining that the quota for women changed some variables and did not change others, how should they interpret it? How will they know that these differences are not due to pure chance rather than the policy? It is necessary to present them with a clear hypothesis of how quotas are supposed to change policymaking, which will help you make predictions about which outcomes are affected.

DISCUSSION TOPIC 2, CONTINUED

3. What might be some examples of key hypotheses you could test? Pick one.

4. Which indicators or combinations of indicators would you use to test your key hypothesis?

Use a logical framework to delineate intermediate and final outcomes

A good way of figuring out the important outcomes is to lay out your theory of change; that is, to draw a logical framework linking the intervention, step by step, to the key final outcomes.

DISCUSSION TOPIC 2, CONTINUED

5. What are the steps or conditions that link quotas (the intervention) to the final outcomes?
6. Which indicators should you try to measure at each step in your logical framework?
7. Using the outcomes and conditions, draw a possible logical framework, linking the intervention and the final outcomes.

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 2: Learn to Read Evaluations

Why Randomize?

This case study is based on "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in India," by Abhijit Banerjee (MIT), Rukmini Banerjee (Pratham), Esther Duflo (MIT), Rachel Glennerster (J-PAL), and Stuti Khemani (The World Bank)

J-PAL thanks the authors for allowing us to use their paper

Key Vocabulary

Counterfactual: what would have happened to the participants in a program had they not received the intervention. The counterfactual cannot be observed from the treatment group; it can only be inferred from the comparison group.

Comparison Group: in an experimental design, a randomly assigned group from the same population that does not receive the intervention, but is the subject of evaluation. Participants in the comparison group are used as a standard for comparison against the treated subjects in order to validate the results of the intervention.

Program Impact: estimated by measuring the difference in outcomes between comparison and treatment groups. The true impact of the program is the difference in outcomes between the treatment group and its counterfactual.

Baseline: data describing the characteristics of participants measured across both treatment and comparison groups prior to implementation of intervention.

Endline: data describing the characteristics of participants measured across both treatment and comparison groups after implementation of intervention.

Selection Bias: statistical bias between comparison and treatment groups in which individuals in one group are systematically different from those in the other. These can occur when the treatment and comparison groups are chosen in a non-random fashion so that they differ from each other by one or more factors that may affect the outcome of the study.

Omitted Variable Bias: statistical bias that occurs when certain variables/characteristics (often unobservable), which affect the measured outcome, are omitted from a regression analysis. Because they are not included as controls in the regression, one incorrectly attributes the measured impact solely to the program.

Introduction

In a large-scale survey conducted in 2004, Pratham discovered that only 39% of children (aged 7-14) in rural Uttar Pradesh could read and understand a simple story, and nearly 15% could not recognize even a letter.

During this period, Pratham was developing the "Learn-to-Read" (L2R) module of its Read India campaign. L2R included a unique pedagogy teaching basic literacy skills, combined with a grassroots organizing effort to recruit volunteers willing to teach.

This program allowed the community to get involved in children's education more directly through village meetings where Pratham staff shared information on the status of literacy in the village and the rights of children to education. In these meetings, Pratham identified community members who were willing to teach. Volunteers attended a training session on the pedagogy, after which they could hold after-school reading classes for children, using materials designed and provided by Pratham. Pratham staff paid occasional visits to these camps to ensure that the classes were being held and to provide additional training as necessary.

Did this program work? How would you measure the impact?

Did the Learn to Read Project work?

Did Pratham's "Learn to Read" program work? What is required in order for us to measure whether a program worked, or whether it had impact?

In general, to ask if a program works is to ask if the program achieves its goal of changing certain outcomes for its participants, and ensure that those changes are not caused by some other factors or events happening at the same time. To show that the program causes the observed changes, we need to simultaneously show that if the program had not been implemented, the observed changes would not have occurred (or would be different). But how do we know what would have happened? If the program happened, it happened. Measuring what would have happened in the absence of the program requires entering an imaginary world in which the program was never given to these participants. The outcomes of the same participants in this imaginary world are referred to as the counterfactual. Since we cannot observe the true counterfactual, the best we can do is to estimate it by mimicking it.

The key challenge of program impact evaluation is constructing or mimicking the counterfactual. We typically do this by selecting a group of people that resemble the participants as much as possible but who did not participate in the program. This group is called the comparison group. Because we want to be able to say that it was the program and not some other factor that caused the changes in outcomes, it is important that the only difference between the comparison group and the participants is that the comparison group did not participate in the program. We then estimate "impact" as the difference observed at the end of the program between the outcomes of the comparison group and the outcomes of the program participants.

The impact estimate is only as accurate as the comparison group is successful at mimicking the counterfactual. If the comparison group poorly represents the counterfactual, the impact is (in most circumstances) poorly estimated. Therefore the method used to select the comparison group is a key decision in the design of any impact evaluation.

That brings us back to our questions: Did the Learn to Read project work? What was its impact on children's reading levels?

In our case, the intention of the program is to "improve children's reading levels" and the reading level is the outcome measure. So, when we ask if the Learn to Read project worked, we are asking if it improved children's reading levels. The impact is the difference between reading levels after the children have taken the reading classes and what their reading level would have been if the reading classes had never existed.

For reference, Reading Level is an indicator variable that takes value 0 if the child can read nothing, 1 if he knows the alphabet, 2 if he can recognize words, 3 if he can read a paragraph, and 4 if he can read a full story.

What comparison groups can we use? The following experts illustrate different methods of evaluating impact. (Refer to the table on the last page of the case for a list of different evaluation methods).

Estimating the impact of the Learn to Read project

METHOD 1:

News Release: Read India helps children Learn to Read.

Pratham celebrates the success of its "Learn to Read" program—part of the Read India Initiative. It has made significant progress in its goal of improving children's literacy rates through better learning materials, pedagogical methods, and most importantly, committed volunteers. The achievement of the "Learn to Read" (L2R) program demonstrates that a revised curriculum, galvanized by community mobilization, can produce significant gains. Massive government expenditures in mid-day meals and school construction have failed to achieve similar results. In less than a year, the reading levels of children who enrolled in the L2R camps improved considerably.

FIGURE 1

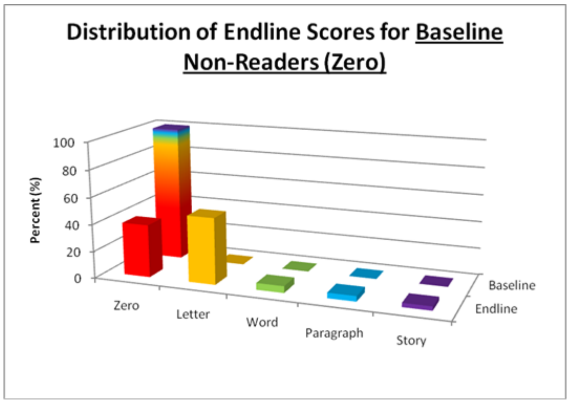
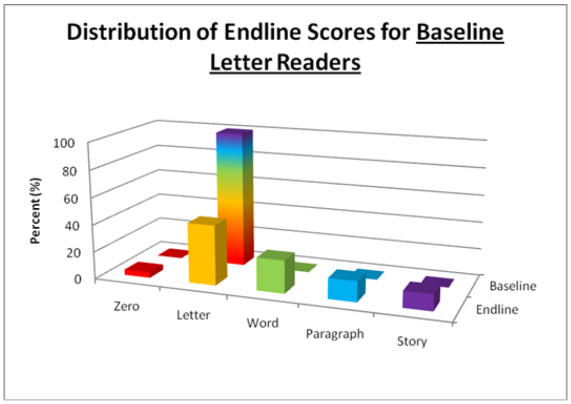


FIGURE 2



Just before the program started, half these children could not recognize Hindi words—many nothing at all. But after spending just a few months in Pratham reading classes, more than half improved by at least one reading level, with a significant number capable of recognizing words and several able to read full paragraphs and stories! *On average, the literacy measure of these students improved by nearly one full reading level during this period.*

DISCUSSION TOPIC 1

Identifying evaluation

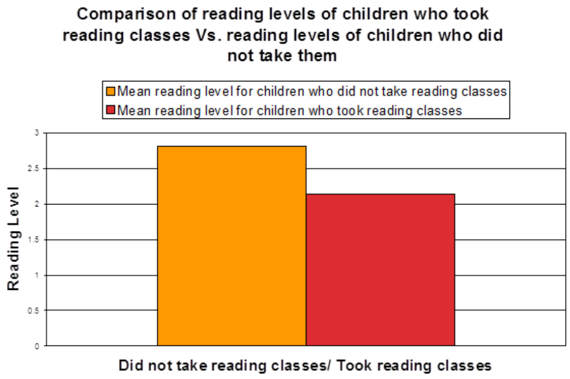
1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 2:

Opinion: The "Read India" project not up to the mark

Pratham has raised millions of dollars, expanding rapidly to cover all of India with its so-called "Learn-to-Read" program, but do its students actually learn to read? Recent evidence suggests otherwise. A team of evaluators from Education for All found that children who took the reading classes ended up with literacy levels significantly below those of their village counterparts. After one year of Pratham reading classes, Pratham students could only recognize words whereas those who steered clear of Pratham programs were able to read full paragraphs.

FIGURE 3



Notes: Reading Level is an indicator variable that takes value 0 if the child can read nothing, 1 if he knows the alphabet, 2 if he can recognize words, 3 if he can read a paragraph and 4 if he can read a full story.

If you have a dime to spare, and want to contribute to the education of India's illiterate children, you may think twice before throwing it into the fountain of Pratham's promises.

DISCUSSION TOPIC 2

Identifying evaluation

1. What type of evaluation does this opinion piece imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 3:

Letter to the Editor: EFA should consider Evaluating Fairly and Accurately

There have been several unfair reports in the press concerning programs implemented by the NGO Pratham. A recent article by a former Education for All bureaucrat claims that Pratham is actually hurting the children it recruits into its 'Learn-to-Read' camps. However, the EFA analysis uses the wrong metric to measure impact. It compares the reading *levels* of Pratham students with other children in the village—not taking into account the fact that Pratham targets those whose literacy levels are particularly poor at the beginning. If Pratham simply recruited the most literate children into their programs, and compared them to their poorer counterparts, they could claim success without conducting a single class. But Pratham does not do this. And realistically, Pratham does not expect its illiterate children to overtake the stronger students in the village. It simply tries to initiate improvement over the current state. Therefore the metric should be *improvement* in reading levels—not the final level. When we repeated EFA's analysis using the more-appropriate outcome measure, the Pratham kids improved at twice the rate of the non-Pratham kids (0.6 reading level increase compared to 0.3). This difference is statistically very significant.

Had the EFA evaluators thought to look at the more appropriate outcome, they would recognize the incredible success of Read India. Perhaps they should enroll in some Pratham classes themselves.

DISCUSSION TOPIC 3

Identifying evaluation

1. What type of evaluation does this letter imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 4:

The numbers don't lie, unless your statisticians are asleep

Pratham celebrates victory, opponents cry foul. A closer look shows that, as usual, the truth is somewhere in between.

There has been a war in the press between Pratham's supporters and detractors. Pratham and its advocates assert that the Read India campaign has resulted in large increases in child literacy. Several detractors claim that Pratham programs, by pulling attention away from the schools, are in fact causing significant harm to the students. Unfortunately, this battle is being waged using instruments of analysis that are seriously flawed. The ultimate victim is the public who is looking for an answer to the question: is Pratham helping its intended beneficiaries?

This report uses sophisticated statistical methods to measure the true impact of Pratham programs. We were concerned about other variables confounding previous results. We therefore conducted a survey in these villages to collect information on child age, grade-level, and parents' education level, and used those to predict child test scores.

Looking at Table 1, we find some positive results, some negative results and some "no-results", depending on which variables we control for. The results from column (1) suggest that Pratham's program hurt the children. There is a negative correlation between receiving Pratham classes and final reading outcomes (-0.68). Column (3), which evaluates improvement, suggests impressive results (0.24). But looking at child outcomes (either level or improvement) controlling for initial reading levels, age, gender, standard and parent's education level – all determinants of child reading levels – we found no impact of Pratham programs.

Therefore, controlling for the right variables, we have discovered that on one hand, Pratham has not caused the harm claimed by certain opponents, but on the other hand, it has not helped children learn. Pratham has therefore failed in its effort to convince us that it can spend donor money effectively.

DISCUSSION TOPIC 4

Identifying evaluation

1. What type of evaluation does this report imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation.

Table 1: Reading Outcomes

Dependent variables: reading level and improvement in reading level are the primary outcomes in this analysis.

Key independent variable: reading classes are the treatment; the analysis tests the effect of these classes on reading outcomes.

Control variables: (independent) variables other than the reading classes that may influence children's reading outcomes

Statistical significance: the corresponding result is unlikely to have occurred by chance, and thus is statistically significant (credible)

	Level		Improvement	
	(1)	(2)	(3)	(4)
Reading Classes	-.68** (0.0829)	0.04 (0.1031)	0.24** (0.0628)	0.11 (0.1081)
Previous Reading Level		0.71** (0.0215)		
Age		0.00 (0.0182)		-0.01 (0.0194)
Sex		-0.01 (0.0469)		0.05 (0.0514)
Standard		0.02 (0.0174)		-0.08** (0.0171)
Parents Literate		0.04 (0.0457)		0.13** (0.0506)
Constant	2.82 (0.0239)	0.36 (0.2648)	0.37 (0.0157)	.75 (0.3293)
School-type controls	No	Yes	No	Yes

Notes: The omitted category for school type is 'Did not go to school." Reading Level is an indicator variable that takes value 0 if the child can read nothing, 1 if he knows the alphabet, 2 if he can recognize words, 3 if he can read a paragraph and 4 if he can read a full story.

	Methodology	Description	Who is in the comparison group?	Required Assumptions	Required Data
Quasi-Experimental Methods	Pre-Post	Measure how program participants improved (or changed) over time.	Program participants themselves—before participating in the program.	The program was the only factor influencing any changes in the measured outcome over time.	Before and after data for program participants.
	Simple Difference of Means	Measure difference between program participants and non-participants after the program is completed.	Individuals who didn't participate in the program (for any reason), but for whom data were collected after the program.	Non-participants are identical to participants except for program participation, and were equally likely to enter program before it started.	After data for program participants and non-participants.
	Differences in Differences	Measure improvement (change) over time of program participants *relative to* the improvement (change) of non-participants.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program.	If the program didn't exist, the two groups would have had identical trajectories over this period.	Before and after data for both participants and non-participants.
	Multivariate Regression	Individuals who received treatment are compared with those who did not, and other factors that might explain differences in the outcomes are "controlled" for.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program. In this case data is not comprised of just indicators of outcomes, but other "explanatory" variables as well.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "control variables" for both participants and non-participants.
	Statistical Matching	Individuals in control group are compared to similar individuals in experimental group.	Exact matching: For each participant, at least one non-participant who is identical *on selected characteristics*. Propensity score matching: non-participants who have a mix of characteristics which predict that they would be as likely to participate as participants.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "variables for matching" for both participants and non-participants.
	Regression Discontinuity Design	Individuals are ranked based on specific, measureable criteria. There is some cutoff that determines whether an individual is eligible to participate. Participants are then compared to non-participants and the eligibility criterion is controlled for.	Individuals who are close to the cutoff, but fall on the "wrong" side of that cutoff, and therefore do not get the program.	After controlling for the criteria (and other measures of choice), the remaining differences between individuals directly below and directly above the cut-off score are not statistically significant and will not bias the results. A necessary but sufficient requirement for this to hold is that the cut-off criteria are strictly adhered to.	Outcomes as well as measures on criteria (and any other controls).
	Instrumental Variables	Participation can be predicted by an incidental (almost random) factor, or "instrumental" variable, that is uncorrelated with the outcome, other than the fact that it predicts participation (and participation affects the outcome).	Individuals who, because of this close to random factor, are predicted not to participate and (possibly as a result) did not participate.	If it weren't for the instrumental variable's ability to predict participation, this "instrument" would otherwise have no effect on or be uncorrelated with the outcome.	Outcomes, the "instrument," and other control variables.
Experimental Method	Randomized Evaluation	Experimental method for measuring a causal relationship between two variables.	Participants are randomly assigned to the control groups.	Randomization "worked." That is, the two groups are statistically identical (on observed and unobserved factors).	Outcome data for control and experimental groups. Control variables can help absorb variance and improve "power".

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 3: Extra Teacher Program

How to Randomize

This case study is based on the paper "Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," by Esther Duflo (MIT), Pascaline Dupas (UCLA), and Michael Kremer (Harvard)

J-PAL thanks the authors for allowing us to use their paper

Key vocabulary

Level of randomization: the level of observation (e.g., individual, household, school, village) at which treatment and comparison groups are randomly assigned.

Introduction

Confronted with overcrowded schools and a shortage of teachers, in 2005 the NGO International Child Support Africa (ICS) offered to help the school system of Western Kenya by introducing contract teachers in 120 primary schools. Under its two-year program, ICS provided funds to these schools to hire one extra teacher per school. In contrast to the civil servants hired by the Ministry of Education, contract teachers are hired locally by school committees. ICS expected this program to improve student learning by, among other things, decreasing class size and using teachers who are more directly accountable to the communities they serve. However, contract teachers tend to have less training and receive a lower monthly salary than their civil servant counterparts. Thus there was concern about whether these teachers were sufficiently motivated, given their compensation, or qualified, given their credentials.

What experimental designs could test the impact of this intervention on educational achievement? Which of these changes in the school landscape is primarily responsible for improved student performance?

Overcrowded schools

Like many other developing countries, Kenya has recently made rapid progress toward the Millennium Development Goal of universal primary education. Largely due to the elimination of school fees in 2003, primary school enrollment rose nearly 30 percent, from 5.9 million to 7.6 million between 2002 and 2005.

Without accompanying government funding, however, this progress has created its own set of new challenges in Kenya:

1. **Large class sizes:** Due to budget constraints, the rise in primary school enrollment has not been matched by proportional increases in the number of teachers. (Teacher salaries already account for the largest component of educational spending.) The result has been very large class sizes, particularly in lower grades. In a sample of schools in Western Kenya, for example, the average first grade class in 2005 had 83 students. This is concerning because it is believed that small classes are most important for the youngest students, who are still acclimating to the school environment. The Kenyan National Union of Teachers estimates that the country needs an additional 60,000 primary school teachers in addition to the existing 175,000 in order to reach all primary students and decrease class sizes.

2. **Teacher absenteeism:** Further exacerbating the problem of high pupil-teacher ratios, teacher absenteeism remains high, reaching nearly 20 percent in some areas of Kenya.

   There are typically no substitutes for absent teachers, so students simply mill around, go home, or join another class, often in a different grade. Small schools, which are prevalent in rural areas of developing countries, may be closed entirely as a result of teacher absence. Families have to consider whether school will even be open when deciding whether or not to send their children to school. An obvious result is low student attendance—even on days when the school is open.

3. **Heterogeneous classes:** Classes in Kenya are also very heterogeneous, with students varying widely in terms of school preparedness and support from home.

   Grouping students into classes sorted by ability (known as tracking, or streaming) is controversial among academics and policymakers. On one hand, if teachers find it easier to teach a homogeneous group of students, tracking could improve school effectiveness and test scores. Many argue, on the other hand, that if students learn in part from their peers, tracking could disadvantage low-achieving students while benefiting high-achieving students, thereby exacerbating inequality.

4. **Scarce school materials:** Because of the high costs of educational inputs and the rising number of students, educational resources other than the teacher are stretched, and in some cases up to four students must share one textbook. Additionally, an already overburdened infrastructure deteriorates faster when forced to serve more children.

5. **Low completion rates:** As a result of these factors, completion rates are very low in Kenya, with only 45.1 percent of boys and 43.3 percent of girls completing the first grade.

All in all, these issues pose a new challenge to the community: how to ensure minimum quality of education given Kenya's budget constraints.

What are contract teachers?

Governments in several developing countries have responded to similar challenges by staffing unfilled teaching positions with locally hired contract teachers who are not civil service employees. There are four main characteristics of contract teachers: they are (1) appointed on annual renewable contracts, with no guarantee of renewed employment (unlike regular civil service teachers); (2) often less qualified than regular teachers and much less likely to have a formal teacher training certificate or degree; (3) paid lower salaries than those of regular teachers (typically less than a fifth of the salaries paid to regular teachers); and (4) more likely to be from the local area where the school is located.

Are contract teachers effective?

The increasing use of contract teachers has been one of the most significant policy innovations in providing primary education in developing countries, but it has also been highly controversial. Supporters say that using contract teachers is an efficient way of expanding education access and quality to a large number of first-generation learners. Knowing that the school committee's decision of whether or not to rehire them the following year may hinge on performance, contract teachers are motivated to try harder than their tenured government counterparts. Contract teachers are also often more similar to their students geographically, culturally, and socioeconomically.

Opponents argue that using underqualified and untrained teachers may staff classrooms, but will not produce learning outcomes. Furthermore, the use of contract teachers de-professionalizes teaching, reduces the prestige of the entire profession, and reduces motivation of all teachers. Even if it helps in the short term, it may hurt efforts to recruit highly qualified teachers in the future.

While the use of contract teachers has generated much controversy, there is very little rigorous evidence regarding the effectiveness of contract teachers in improving student learning outcomes.

The Extra Teacher Program randomized evaluation

In January 2005, ICS Africa initiated a two-year program to examine the effect of contract teachers on education in Kenya. Under the program, ICS gave funds to 120 local school committees to hire one extra contract teacher to teach an additional first grade class. The purpose of this intervention was to address three challenges: class size, teacher accountability, and heterogeneity of ability. The evaluation was designed to measure the impact of class-size reductions, the relative effectiveness of contract teachers, and how tracking by ability would impact both low- and high-achieving students.

Addressing multiple research questions through experimental design

Different randomization strategies may be used to answer different questions. What strategies could be used to evaluate the following questions? How would you design the study? Who would be in the treatment and control groups, and how would they be randomly assigned to these groups?

DISCUSSION TOPIC 1

Testing the effectiveness of contract teachers

1. What is the relative effectiveness of contract teachers versus regular government teachers?

DISCUSSION TOPIC 2

Looking at more general approaches to improving education

1. What is the effect of grouping students by ability on student performance?
2. What is the effect of smaller class sizes on student performance?

DISCUSSION TOPIC 3

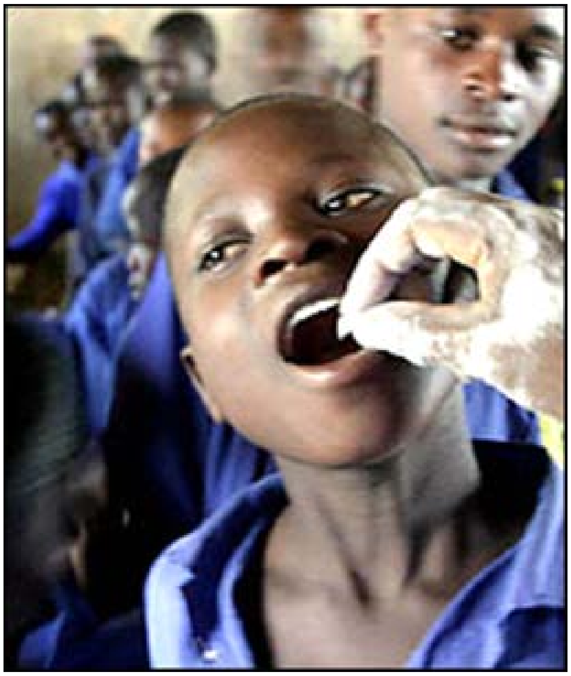
Addressing all questions with a single evaluation

1. Could a single evaluation explore all of these issues at once?
2. What randomization strategy could do so?

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 4: Deworming in Kenya

Addressing threats to experimental integrity



This case study is based on Edward Miguel and Michael Kremer, “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities,” Econometrica 72(1): 159-217, 2004.

J-PAL thanks the authors for allowing us to use their paper.

Key vocabulary

Phase-in design: a study design in which groups are individually phased into treatment over a period of time; groups that are scheduled to receive treatment later act as the comparison groups in earlier rounds.

Equivalence: when groups are identical on all baseline characteristics, both observable and unobservable. It is ensured by randomization, *in expectation*.

Attrition: the process of individuals dropping out of either the treatment or comparison group over the course of the study.

Attrition bias: a statistical bias that occurs when individuals systematically drop out of either the treatment or the comparison group for reasons related to the treatment.

Partial compliance: when individuals do not "comply" with their assignment (to treatment or comparison). Also termed "diffusion" or "contamination."

Intention to Treat: measured impact of a program that compares outcomes from all individuals assigned to the treatment group to those assigned to the control group (regardless of whether they actually availed the treatment). Often termed, "Average Treatment Effect" (ATE)

Local Average Treatment Effect (LATE): the estimated impact of a program on participants who participated in the program solely because they were assigned to the treatment group (also called the Complier Average Causal Effect or CACE). This is different from the Intention to Treat estimate when there is partial compliance. A special case of the Local Average Treatment Effect, called the Treatment on the Treated estimate, occurs when partial compliance only occurs in the Treatment Group.

Externality: an indirect cost or benefit incurred by individuals who did not directly receive the treatment. Also known as a "spillover".

Introduction

Between 1998 and 2001, the NGO International Child Support Africa (ICS) implemented a school-based mass deworming program in 75 primary schools in western Kenya. The program treated the 45,000 pupils enrolled at these schools for worms—hookworm, roundworm, whipworm, and schistosomiasis. Schools were phased in randomly.

Randomization ensures that the treatment and comparison groups are comparable at the beginning, but there can be external influences that can make them incomparable at the end of the program. Imagine we have a pile of seeds from five different plants. If we split this pile randomly into two bags, both bags should have the same composition of seeds. Suppose now that one of the bags gets punctured; the hole is small enough for only the smallest seed variety to pass through. What can we say about the composition of the two bags after this event? Are the two bags still comparable? This type of event can happen between initial randomization and the endline and can reintroduce selection bias; it diminishes the validity of the impact estimates and is a threat to the integrity of the experiment.

How can common threats to experimental integrity be managed?

Worms: a common problem with a cheap solution

Worm infections account for over 40 percent of the global tropical disease burden. Infections are common in areas with poor sanitation. More than 2 billion people are affected. Children, who typically have poorer sanitary habits, are particularly vulnerable: 400 million school-age children are chronically infected with intestinal worms.

Symptoms include listlessness, diarrhea, abdominal pain, and anemia. But worms affect more than the health of children. Heavy worm infections can impair children's physical and mental development, leading to poor attendance and performance in school.

Poor sanitation and personal hygiene habits facilitate transmission. Infected people excrete worm eggs in their feces and urine. In areas with poor sanitation, the eggs contaminate the soil or water. Other people are infected when they ingest contaminated food or soil (hookworm, whipworm, and roundworm), or when hatched worm larvae penetrate their skin upon contact with contaminated soil (hookworm) or fresh water (schistosome). School-age children are more likely to spread worms because they have riskier hygiene practices (more likely to swim in contaminated water, more likely to not use the latrine, less likely to wash hands before eating). So treating a child not only reduces her own worm load; it may also reduce disease transmission—and so benefit the community at large.

Treatment kills worms in the body, but does not prevent reinfection. Oral medication that can kill 99 percent of worms in the body is available: albendazole or mebendazole for treating hookworm, roundworm, and whipworm infections; and praziquantel for treating schistosomiasis. These drugs are cheap and safe. A dose of albendazole or mebendazole costs less than 3 US cents while one dose of praziquantel costs less than 20 US cents. The drugs have very few and minor side effects.

Worms colonize the intestines and the urinary tract, but they do not reproduce in the body; their numbers build up only through repeated contact with contaminated soil or water. The World Health Organization (WHO) recommends presumptive school-based mass deworming in areas with high prevalence. Schools with hookworm, whipworm, and roundworm prevalence over 50 percent should be mass treated with albendazole every 6 months, and schools with schistosomiasis prevalence over 30 percent should be mass treated with praziquantel once a year.

The Primary School Deworming Program

International Child Support Africa (ICS) implemented the Primary School Deworming Program (PSDP) in the Busia District in Western Kenya, a densely settled region with high worm prevalence. Treatment followed WHO guidelines. The medicine was administered by public health nurses from the Ministry of Health, in the presence of health officers from ICS.

The PSDP was expected to affect health, nutrition, and education. To measure impact, ICS collected data on a series of outcomes: prevalence of worm infection, worm loads (severity of worm infection); self-reported illness; and school participation rates and test scores.

Evaluation design: the experiment as planned

Because of administrative and financial constraints, the PSDP could not be implemented in all schools immediately. Instead, the 75 schools were randomly divided into three groups of 25 schools and phased in over three years. Group 1 schools were treated starting in both 1998 and 1999, Group 2 schools in 1999, and Group 3 schools starting in 2001. Group 1 schools were the treatment group in 1998, while schools in Group 2 and Group 3 were the comparison. In 1999 Group 1 and Group 2 schools formed the treatment group and Group 3 schools the comparison.

TABLE 1
The planned experiment: the PSDP treatment timeline showing experimental groups in 1998 and 1999

	1998	1999	2000
Group 1	Treatment	Treatment	Treatment
Group 2	Comparison	Treatment	Treatment
Group 3	Comparison	Comparison	Treatment

For the purpose of the following questions, we will only look at results after the 1998 period.

Threats to integrity of the planned experiment

DISCUSSION TOPIC 1
Threats to experimental integrity

Randomization ensures that the groups are in expectation equivalent, and therefore comparable, at the beginning of the program. The impact is then estimated as the difference in the average outcome of the treatment group and the average outcome of the comparison group, both at the end of the program. To be able to say that the program caused the impact, you need to be able to say that the program was the only difference between the treatment and comparison groups over the course of the evaluation.

a. What does it mean to say that the groups are equivalent at the start of the program?

b. Can you check if the groups are equivalent at the beginning of the program? How?

Managing attrition: when the groups do not remain equivalent

Attrition is when people drop out of the sample—both treatment and comparison groups—over the course of the experiment. One common example in clinical trials is when people die; so common indeed that attrition is sometimes called experimental mortality.

DISCUSSION TOPIC 2
Managing attrition

You are looking at the health effects of deworming. In particular you are looking at the worm load (severity of worm infection). Worm loads are scaled as follows:

Heavy worm infections = score of 3

Medium worm infections = score of 2

Light infections = score of 1

There are 30,000 children: 15,000 in treatment schools and 15,000 in comparison schools. After you randomize, the treatment and comparison groups are equivalent, meaning children from each of the three worm load categories are equally represented in both groups.

Suppose protocol compliance is 100 percent: all children who are in the treatment get treated and none of the children in the comparison are treated. Children that were dewormed at the beginning of the school year (that is, children in the treatment group) end up with a worm load of 1 at the end of the year. The number of children in each worm-load category is shown for both the pretest and posttest.

TABLE 2

	Pretest		Posttest	
Worm Load	T	C	T	C
3	5,000	10,000	0	10,000
2	5,000	10,000	0	10,000
1	5,000	10,000	15,000	10,000
Total children tested at school	15,000	30,000	15,000	30,000

Average		

1.

a. At pretest, what is the average worm load for each group?

b. At posttest, what is the average worm load for each group?

c. What is the impact of the program?

d. Do you need to know pretest values? Why or why not?

Suppose now that children who have a worm load of 3 only attend half the time and drop out of school if they are not treated. The number of children in each worm-load category is shown for both the pretest and posttest.

TABLE 3

	Pretest		Posttest	
Worm Load	T	C	T	C
3	5,000	10,000	0	
2	5,000	10,000	0	10,000
1	5,000	10,000	15,000	10,000
Total children tested at school	15,000	30,000	15,000	20,000
Average				

2.

a. At posttest, what is the new average worm load for the comparison group?

b. What is the impact of the program?

c. Is this outcome difference an accurate estimate of the impact of the program? Why or why not?

d. If it is not accurate, does it overestimate or underestimate the impact?

e. How can we get a better estimate of the program's impact?

Besides worm load, the PSDP also looked at outcome measures such as school attendance rates and test scores.

3.

a. Would differential attrition (i.e., differences in dropouts between treatment and comparison groups) bias either of these outcomes? How?

b. Would the impacts on these final outcome measures be underestimated or overestimated?

In Case Study 1, you learned about other methods to estimate program impact, such as pre-post, simple difference, difference-in-difference, and multivariate regression.

4.

a. Does the threat of attrition only present itself in randomized evaluations?

Managing partial compliance: when the treatment group does not actually get treated or the comparison group gets treated

Some people assigned to the treatment may in the end not actually get treated. In an after-school tutoring program, for example, some children assigned to receive tutoring may simply not show up for tutoring. Those assigned to the comparison group may obtain access to tutoring, either from the program or from another provider. Or comparison group children may get extra help from the teachers or acquire program materials and methods from their classmates. In any of these scenarios, people are not complying with their assignment in the planned experiment. This is called "partial compliance" or "diffusion" or, less benignly, "contamination." In contrast to carefully controlled lab experiments, diffusion is a ubiquitous concern in social programs. After all, life goes on, people will be people, and you have no control over what they decide to do over the course of the experiment. All you can do is plan your experiment and offer them treatments. How, then, can you deal with the complications that arise from partial compliance?

DISCUSSION TOPIC 3

Managing partial compliance

TABLE 4

	Pretest		Posttest	
Worm Load	T	C	T	C
3	5,000	10,000	5,000	10,000
2	5,000	10,000	0	10,000
1	5,000	10,000	10,000	10,000
Total children tested at school	15,000	30,000	15,000	30,000

1.

a. Calculate the impact estimate based on the original group assignments.

b. This is an unbiased measure of the effect of the program, but in what ways is it useful and in what ways is it not as useful?

c. Five of your colleagues are passing by your desk; they all agree that you should calculate the effect of the treatment using only the 10,000 children who were treated and compare them to the comparison group. Is this advice sound? Why or why not?

d. Another colleague says that it's not a good idea to drop the untreated entirely; you should use them but consider them as part of the comparison. Is this advice sound? Why or why not?

Managing spillovers: when the comparison, itself untreated, benefits from the treatment being treated

People assigned to the control group may benefit indirectly from those receiving treatment. For example, a program that distributes insecticide-treated nets may reduce malaria transmission in the community, indirectly benefiting those who themselves do not sleep under a net. Such effects are called externalities or spillovers.

DISCUSSION TOPIC 4

Managing spillovers

In the deworming program, randomization was at the school level. However, while all boys at a given treatment school were treated, only girls younger than thirteen received the deworming pill. This was due to the fact that the WHO had not tested (and thus not yet approved) the deworming pill for pregnant women. Because it was difficult to determine which girls were at risk of getting pregnant, the program decided to not administer the medication to any girl thirteen or older. (Postscript: since the deworming evaluation was implemented, the WHO has approved the deworming medication for pregnant women.)

Thus, at a given treatment school, there was a distinct group of students that was never treated, but lived in very close proximity to a group that was treated.

Suppose protocol compliance is 100 percent: all boys and girls under thirteen in treatment schools get treated and all girls thirteen and over in treatment schools, as well as all children in comparison schools, do not get treated.

You can assume that due to proper randomization, the distribution of worm load across the three groups of students is equivalent between treatment and control schools prior to the intervention.

TABLE 5

	Posttest					
	Treatment			Comparison		
Worm Load	All boys	Girls <13 yrs	Girls >= 13 yrs	All boys	Girls <13 yrs	Girls >= 13 yrs
3	0	0	0	5000	2000	2000
2	0	0	2000	5000	3000	3000
1	10000	5000	3000	0	0	0
Total children tested at school	20000			20000		

1.

a. If there are any spillovers, where would you expect them to show up?

b. Is it possible for you to capture these potential spillover effects? How?

2.

a. What is the treatment effect for boys in treatment versus comparison schools?

b. What is the treatment effect for girls under thirteen in treatment versus comparison schools?

c. What is the direct treatment effect among those who were treated?

d. What is the treatment effect for girls thirteen and older in treatment versus comparison schools?

e. What is the indirect treatment effect due to spillovers?

f. What is the total program effect?

Exercise A: Understanding random sampling and the law of large numbers

In this exercise, we will visually explore random samples of different sizes from a given population. In particular, we will try to demonstrate that larger sample sizes tend to be more reflective of the underlying population.

1. Open the file "Exercise A_SamplingDistributions.xlsm".
2. If prompted, select "Enable Macros".
3. Navigate to the "Randomize" worksheet, which allows you to choose a random sample of size "Sample Size" from the data contained in the "control" worksheet.
4. Enter "10" for "Sample Size and click the "Randomize" button. Observe the distribution of the various characteristics between Treatment, Control and Expected. With a sample size this small, the percentage difference from the expected average is quite high for reading scores. Click "Randomize" multiple times and observe how the distribution changes.
5. Now, try "50" for the sample size. What happens to the distributions? Randomize a few times and observe the percentage difference for the reading scores.
6. Increase the sample size to "500", "2000" and "10000", and repeat the observations from step 5. What can we say about larger sample sizes? How do they affect our Treatment and Control samples? Should the percentage difference between Treatment, Control and Expected always go down as we increase sample size?

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Exercise B: How to do Random Assignment using MS Excel

CONTENTS

Intro

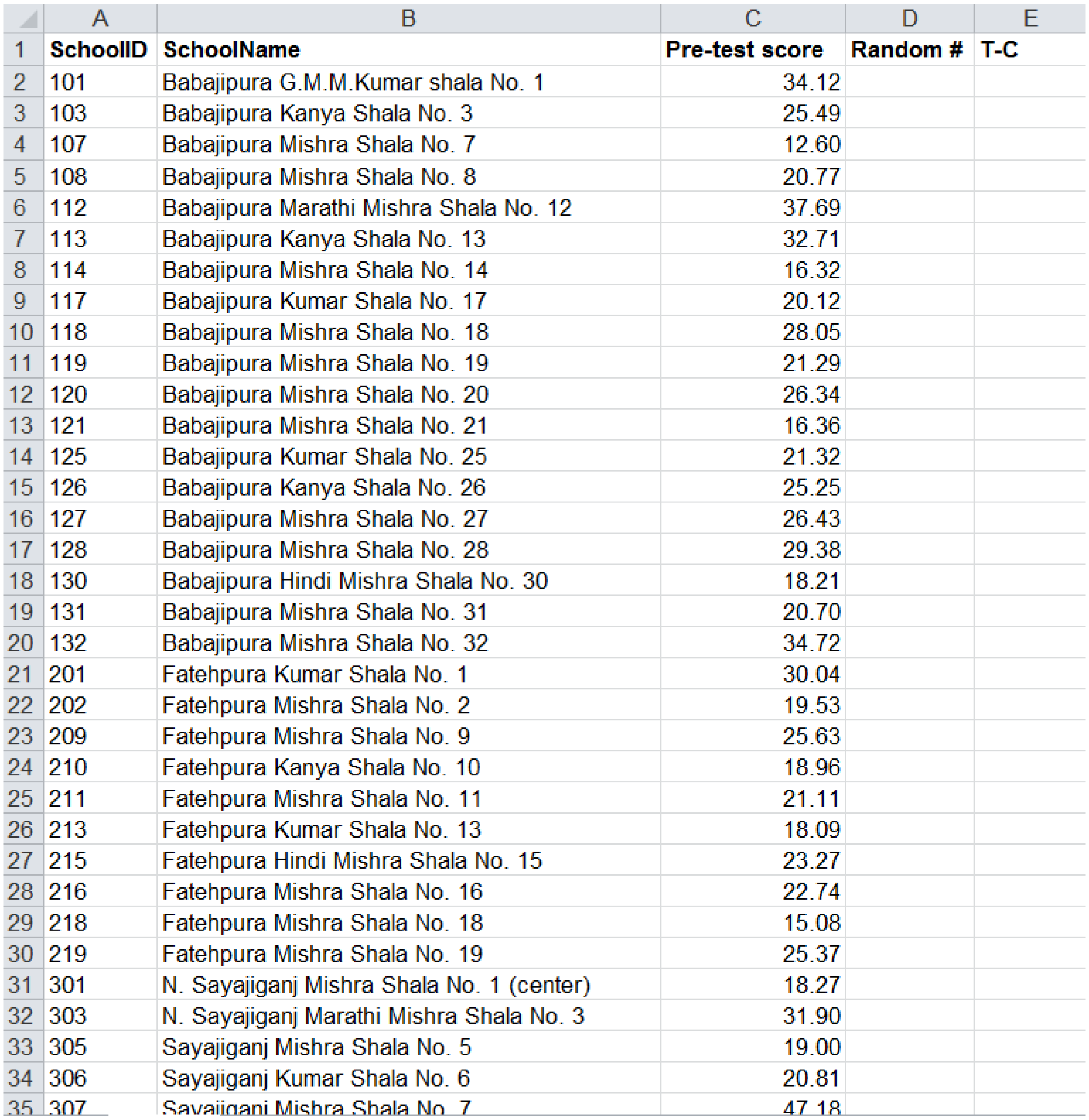
Like most spreadsheet programs, MS Excel can generate random numbers on command. MS Excel has two native random-number-generating functions. The first, =RAND(), creates a *continuous* random number between 0 and 1—it could be any number of 9 decimal places between 0 and 1. The second, =RANDBETWEEN(*bottom*, *top*) creates *integers* between any two integer values within a range, where you specify the *bottom* and *top* of that range.

Part 1: Simple Randomization

Say we had a list of schools and we wanted to assign them to treatment or control based on a coin flip (heads = treatment and tails = control). We can do this by randomly generating the value of 0 or 1 using the RANDBETWEEN function, and choosing 0 and 1 as the range. We could then assign all schools with 0 to the control group, and all schools with 1 to the treatment group (or vice versa). This is equivalent to a coin flip where 0 represents tails and 1 represents control. Equivalently, we could produce a *continuous* random number for each observation and assign those with (say) random number greater than or equal to 0.5 to treatment and smaller than 0.5 to control.

The illustration below shows how to do this step-by-step.

We have a list of all schools



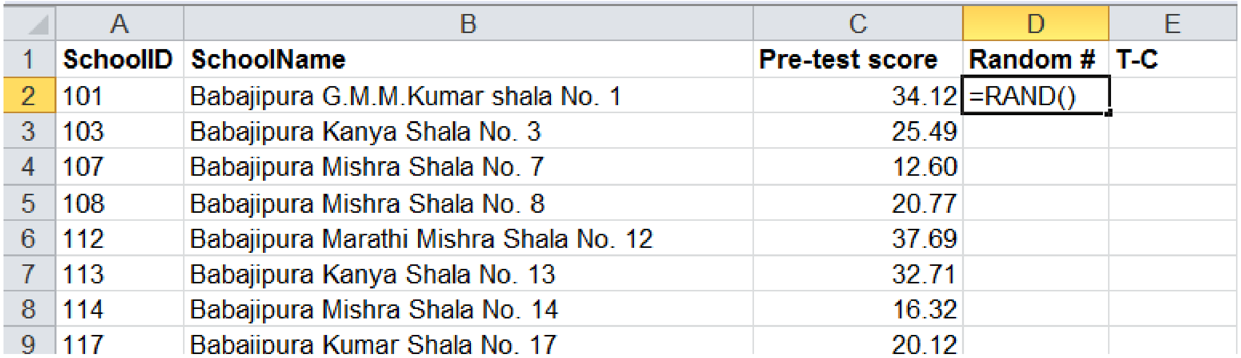
	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	34.12		
3	103	Babajipura Kanya Shala No. 3	25.49		
4	107	Babajipura Mishra Shala No. 7	12.60		
5	108	Babajipura Mishra Shala No. 8	20.77		
6	112	Babajipura Marathi Mishra Shala No. 12	37.69		
7	113	Babajipura Kanya Shala No. 13	32.71		
8	114	Babajipura Mishra Shala No. 14	16.32		
9	117	Babajipura Kumar Shala No. 17	20.12		
10	118	Babajipura Mishra Shala No. 18	28.05		
11	119	Babajipura Mishra Shala No. 19	21.29		
12	120	Babajipura Mishra Shala No. 20	26.34		
13	121	Babajipura Mishra Shala No. 21	16.36		
14	125	Babajipura Kumar Shala No. 25	21.32		
15	126	Babajipura Kanya Shala No. 26	25.25		
16	127	Babajipura Mishra Shala No. 27	26.43		
17	128	Babajipura Mishra Shala No. 28	29.38		
18	130	Babajipura Hindi Mishra Shala No. 30	18.21		
19	131	Babajipura Mishra Shala No. 31	20.70		
20	132	Babajipura Mishra Shala No. 32	34.72		
21	201	Fatehpura Kumar Shala No. 1	30.04		
22	202	Fatehpura Mishra Shala No. 2	19.53		
23	209	Fatehpura Mishra Shala No. 9	25.63		
24	210	Fatehpura Kanya Shala No. 10	18.96		
25	211	Fatehpura Mishra Shala No. 11	21.11		
26	213	Fatehpura Kumar Shala No. 13	18.09		
27	215	Fatehpura Hindi Mishra Shala No. 15	23.27		
28	216	Fatehpura Mishra Shala No. 16	22.74		
29	218	Fatehpura Mishra Shala No. 18	15.08		
30	219	Fatehpura Mishra Shala No. 19	25.37		
31	301	N. Sayajiganj Mishra Shala No. 1 (center)	18.27		
32	303	N. Sayajiganj Marathi Mishra Shala No. 3	31.90		
33	305	Sayajiganj Mishra Shala No. 5	19.00		
34	306	Sayajiganj Kumar Shala No. 6	20.81		
35	307	Sayajiganj Mishra Shala No. 7	47.18		

Step 1: Assign a random number to each school

The function RAND () is Excel's basic random number generator. To use it, go to Column D and type

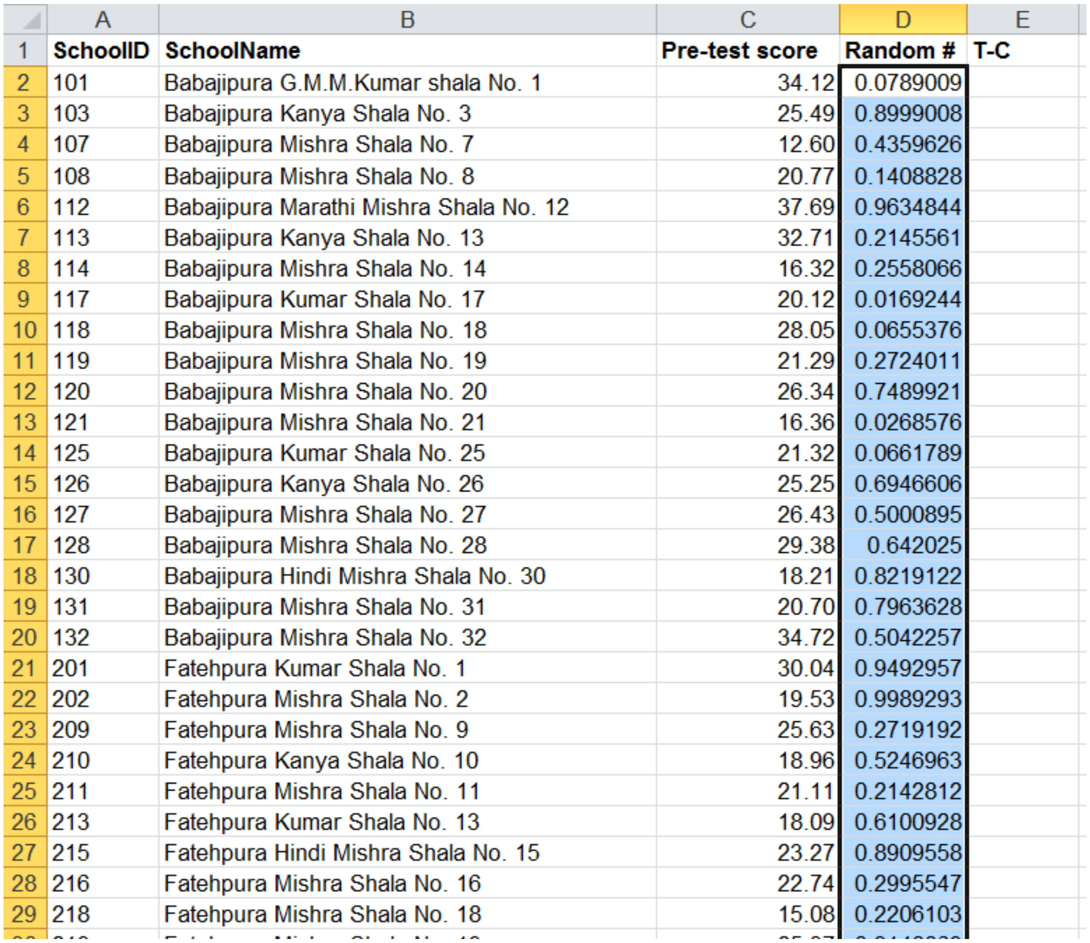
=RAND()

in each cell, adjacent to each school name. Or you can type this function in the top row (row 2) and simply copy and paste to the entire column, or click and drag.



	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	34.12	=RAND()	
3	103	Babajipura Kanya Shala No. 3	25.49		
4	107	Babajipura Mishra Shala No. 7	12.60		
5	108	Babajipura Mishra Shala No. 8	20.77		
6	112	Babajipura Marathi Mishra Shala No. 12	37.69		
7	113	Babajipura Kanya Shala No. 13	32.71		
8	114	Babajipura Mishra Shala No. 14	16.32		
9	117	Babajipura Kumar Shala No. 17	20.12		

Typing =RAND() puts a 9-digit random number between 0 and 1 in the cell.



	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	34.12	0.0789009	
3	103	Babajipura Kanya Shala No. 3	25.49	0.8999008	
4	107	Babajipura Mishra Shala No. 7	12.60	0.4359626	
5	108	Babajipura Mishra Shala No. 8	20.77	0.1408828	
6	112	Babajipura Marathi Mishra Shala No. 12	37.69	0.9634844	
7	113	Babajipura Kanya Shala No. 13	32.71	0.2145561	
8	114	Babajipura Mishra Shala No. 14	16.32	0.2558066	
9	117	Babajipura Kumar Shala No. 17	20.12	0.0169244	
10	118	Babajipura Mishra Shala No. 18	28.05	0.0655376	
11	119	Babajipura Mishra Shala No. 19	21.29	0.2724011	
12	120	Babajipura Mishra Shala No. 20	26.34	0.7489921	
13	121	Babajipura Mishra Shala No. 21	16.36	0.0268576	
14	125	Babajipura Kumar Shala No. 25	21.32	0.0661789	
15	126	Babajipura Kanya Shala No. 26	25.25	0.6946606	
16	127	Babajipura Mishra Shala No. 27	26.43	0.5000895	
17	128	Babajipura Mishra Shala No. 28	29.38	0.642025	
18	130	Babajipura Hindi Mishra Shala No. 30	18.21	0.8219122	
19	131	Babajipura Mishra Shala No. 31	20.70	0.7963628	
20	132	Babajipura Mishra Shala No. 32	34.72	0.5042257	
21	201	Fatehpura Kumar Shala No. 1	30.04	0.9492957	
22	202	Fatehpura Mishra Shala No. 2	19.53	0.9989293	
23	209	Fatehpura Mishra Shala No. 9	25.63	0.2719192	
24	210	Fatehpura Kanya Shala No. 10	18.96	0.5246963	
25	211	Fatehpura Mishra Shala No. 11	21.11	0.2142812	
26	213	Fatehpura Kumar Shala No. 13	18.09	0.6100928	
27	215	Fatehpura Hindi Mishra Shala No. 15	23.27	0.8909558	
28	216	Fatehpura Mishra Shala No. 16	22.74	0.2995547	
29	218	Fatehpura Mishra Shala No. 18	15.08	0.2206103	

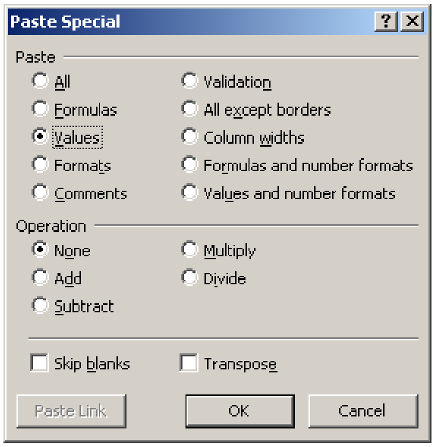
Step 2: Copy the cells in Column D, then paste the values over the same cells

The function =RAND() will re-randomize each time you make any changes to any other part of the spreadsheet. Excel does this because it recalculates all values with any change to any cell. (You can also induce recalculation, and hence re-randomization, by pressing the F9 key.)

Once we've generated our column of random numbers, we do not need to re-randomize. We already have a clean column of random values. To stop Excel from recalculating, you can replace the "functions" in this column with the "values".

To do this, highlight all values in Column D. Then right-click anywhere in the highlighted column, and choose "Copy".

Then, right-click anywhere in that column and choose "Paste Special." The "Paste Special" window will appear. Click on "Values".



Paste

○ All ○ Validation
○ Formulas ○ All except borders
● Values ○ Column widths
○ Formats ○ Formulas and number formats
○ Comments ○ Values and number formats

Operation

● None ○ Multiply
○ Add ○ Divide
○ Subtract

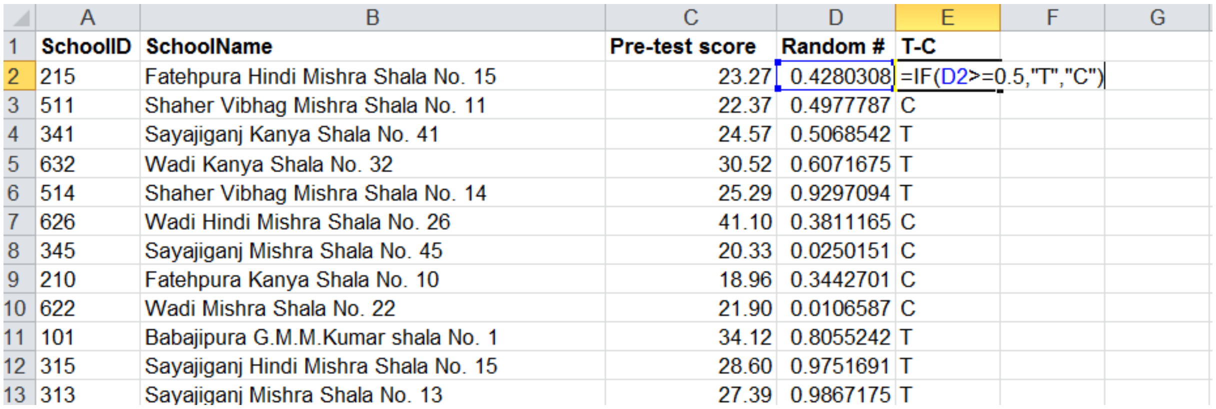
□ Skip blanks □ Transpose

Step 3: Assign treatment/control status for each group

Now use the IF function to assign schools to treatment and control. Go to column E and type

=IF(D2>=0.5,"T","C")

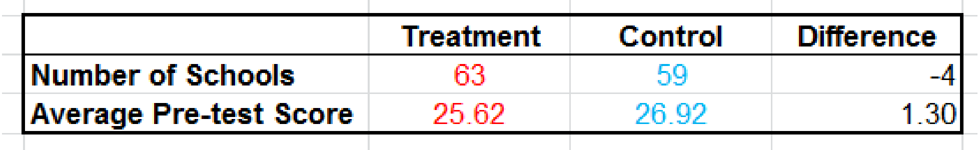
And click and drag (or copy and paste) to the rest of the column. This will enter a "T" for schools that have a random number greater than or equal to 0.5 and "C" for schools with random number less than 0.5.



	A	B	C	D	E	F	G
1	SchoolID	SchoolName	Pre-test score	Random #	T-C		
2	215	Fatehpura Hindi Mishra Shala No. 15	23.27	0.4280308	=IF(D2>=0.5,"T","C")		
3	511	Shaher Vibhag Mishra Shala No. 11	22.37	0.4977787	C		
4	341	Sayajiganj Kanya Shala No. 41	24.57	0.5068542	T		
5	632	Wadi Kanya Shala No. 32	30.52	0.6071675	T		
6	514	Shaher Vibhag Mishra Shala No. 14	25.29	0.9297094	T		
7	626	Wadi Hindi Mishra Shala No. 26	41.10	0.3811165	C		
8	345	Sayajiganj Mishra Shala No. 45	20.33	0.0250151	C		
9	210	Fatehpura Kanya Shala No. 10	18.96	0.3442701	C		
10	622	Wadi Mishra Shala No. 22	21.90	0.0106587	C		
11	101	Babajipura G.M.M.Kumar shala No. 1	34.12	0.8055242	T		
12	315	Sayajiganj Hindi Mishra Shala No. 15	28.60	0.9751691	T		
13	313	Sayajiganj Mishra Shala No. 13	27.39	0.9867175	T		

Your list of schools has now been randomly assigned to treatment and control!

Is the number of schools in in both groups the same? We also have the average pre-test scores for each school. Does the average pre-test score look balanced between the two groups?



	Treatment	Control	Difference
Number of Schools	63	59	-4
Average Pre-test Score	25.62	26.92	1.30

Note, however, that the number of schools in treatment and control will vary each time you re-randomize, as will the average pre-test score. To check this, repeat step 1, but this time instead of copy pasting values, press the F9 key to re-randomize. Re-randomize 10 times and see what happens to the number of schools and the average pre-test score in each group.

Does the number of schools change when you re-randomize? Does the average pre-test score look balanced every time you re-randomize?

Try the above steps using the RANDBETWEEN() function instead of the RAND() function. Do you expect significantly different results? How does the "IF" function change?

Part 2: Complete randomization

Say we had a list of schools and wanted to assign exactly half of them to treatment and half to control

Step 1: Assign a random number to each school

Go to Column D and type:

=RAND()

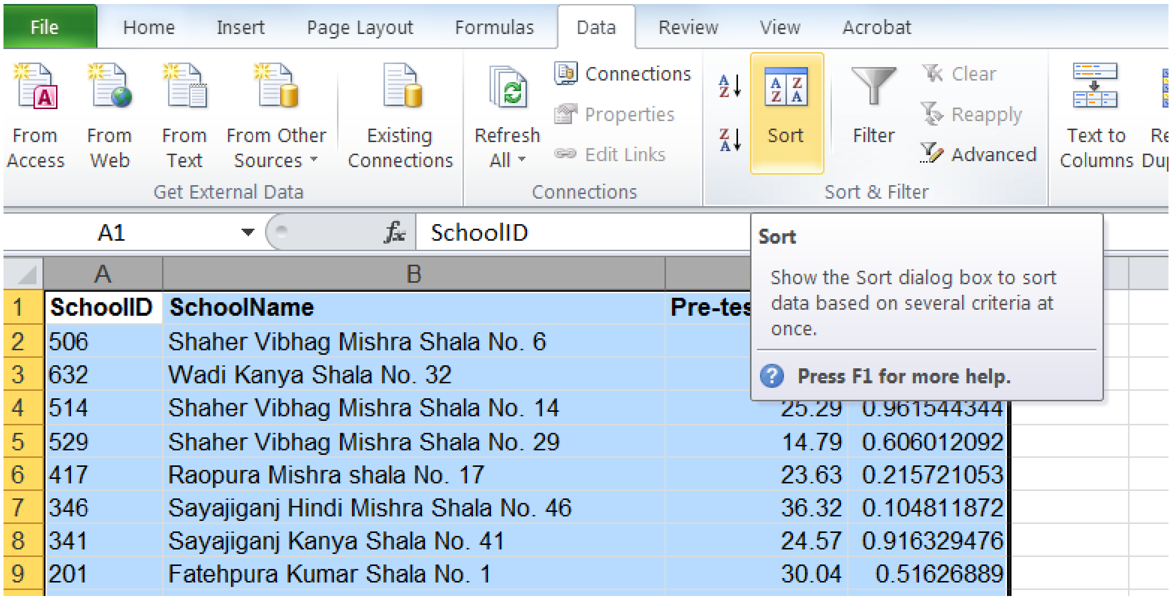
And click and drag (or copy and paste) to the entire column.

Step 2: Copy the cells in Column D, then paste the values over the same cells

Highlight all values in Column D. Then right-click anywhere in the highlighted column, and choose "Copy". Then, right-click anywhere in that column and choose "Paste Special."

Step 3: Sort the columns in either descending or ascending order of Column D

Highlight columns A, B, C and D. In the data tab, press the "Sort" button:



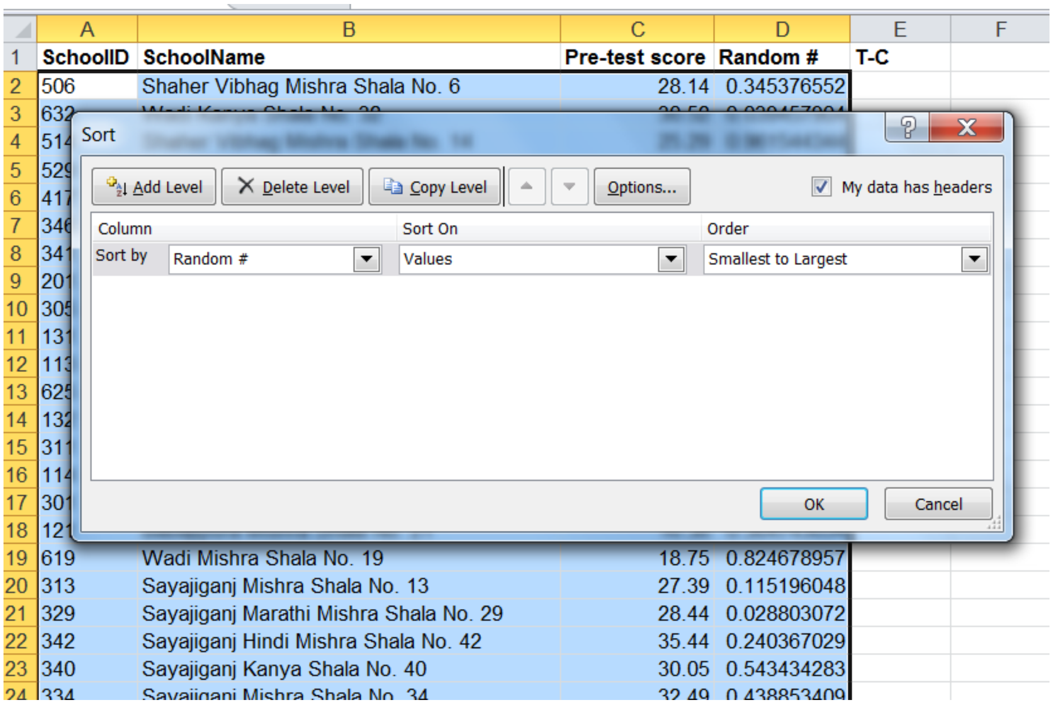
Sort

Show the Sort dialog box to sort data based on several criteria at once.

Press F1 for more help.

	A	B		
1	SchoolID	SchoolName	Pre-tes	
2	506	Shaher Vibhag Mishra Shala No. 6		
3	632	Wadi Kanya Shala No. 32		
4	514	Shaher Vibhag Mishra Shala No. 14	25.29	0.961544344
5	529	Shaher Vibhag Mishra Shala No. 29	14.79	0.606012092
6	417	Raopura Mishra shala No. 17	23.63	0.215721053
7	346	Sayajiganj Hindi Mishra Shala No. 46	36.32	0.104811872
8	341	Sayajiganj Kanya Shala No. 41	24.57	0.916329476
9	201	Fatehpura Kumar Shala No. 1	30.04	0.51626889

A Sort box will pop up.



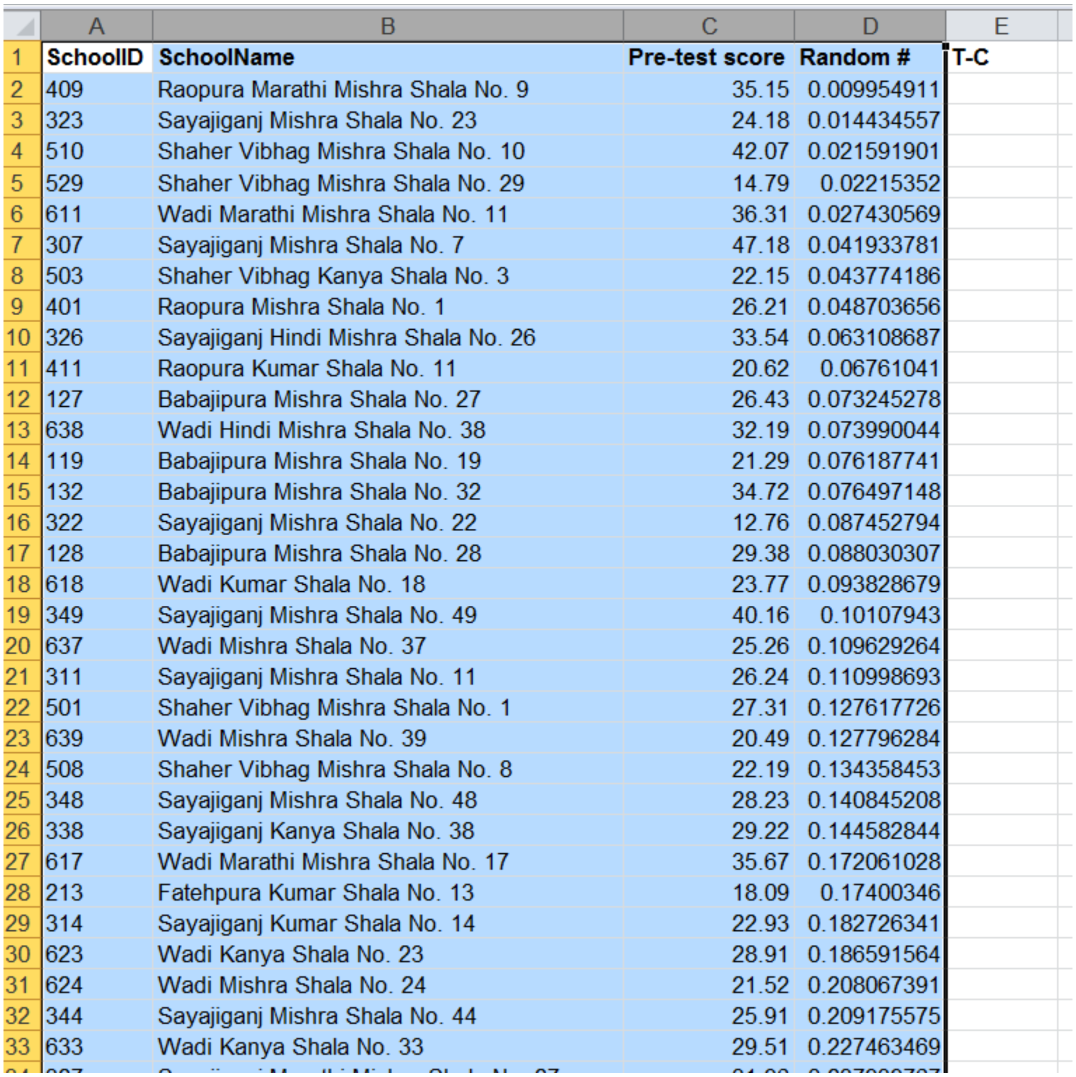
☑ My data has headers

Column	Sort On	Order
Sort by Random #	Values	Smallest to Largest

	A	B	C	D	E	F
1	SchoolID	SchoolName	Pre-test score	Random #	T-C	
2	506	Shaher Vibhag Mishra Shala No. 6	28.14	0.345376552		
3	632					
4	514					
5	529					
6	417					
7	346					
8	341					
9	201					
10	305					
11	131					
12	113					
13	625					
14	132					
15	311					
16	114					
17	301					
18	121					
19	619	Wadi Mishra Shala No. 19	18.75	0.824678957		
20	313	Sayajiganj Mishra Shala No. 13	27.39	0.115196048		
21	329	Sayajiganj Marathi Mishra Shala No. 29	28.44	0.028803072		
22	342	Sayajiganj Hindi Mishra Shala No. 42	35.44	0.240367029		
23	340	Sayajiganj Kanya Shala No. 40	30.05	0.543434283		
24	334	Sayajiganj Mishra Shala No. 34	32.49	0.438853409		

In the “Sort by” column, select “Random #.” Click OK. Doing this sorts the list by the random number in ascending or descending order, whichever you chose.

There! You have a randomly sorted list.



	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	409	Raopura Marathi Mishra Shala No. 9	35.15	0.009954911	
3	323	Sayajiganj Mishra Shala No. 23	24.18	0.014434557	
4	510	Shaher Vibhag Mishra Shala No. 10	42.07	0.021591901	
5	529	Shaher Vibhag Mishra Shala No. 29	14.79	0.02215352	
6	611	Wadi Marathi Mishra Shala No. 11	36.31	0.027430569	
7	307	Sayajiganj Mishra Shala No. 7	47.18	0.041933781	
8	503	Shaher Vibhag Kanya Shala No. 3	22.15	0.043774186	
9	401	Raopura Mishra Shala No. 1	26.21	0.048703656	
10	326	Sayajiganj Hindi Mishra Shala No. 26	33.54	0.063108687	
11	411	Raopura Kumar Shala No. 11	20.62	0.06761041	
12	127	Babajipura Mishra Shala No. 27	26.43	0.073245278	
13	638	Wadi Hindi Mishra Shala No. 38	32.19	0.073990044	
14	119	Babajipura Mishra Shala No. 19	21.29	0.076187741	
15	132	Babajipura Mishra Shala No. 32	34.72	0.076497148	
16	322	Sayajiganj Mishra Shala No. 22	12.76	0.087452794	
17	128	Babajipura Mishra Shala No. 28	29.38	0.088030307	
18	618	Wadi Kumar Shala No. 18	23.77	0.093828679	
19	349	Sayajiganj Mishra Shala No. 49	40.16	0.10107943	
20	637	Wadi Mishra Shala No. 37	25.26	0.109629264	
21	311	Sayajiganj Mishra Shala No. 11	26.24	0.110998693	
22	501	Shaher Vibhag Mishra Shala No. 1	27.31	0.127617726	
23	639	Wadi Mishra Shala No. 39	20.49	0.127796284	
24	508	Shaher Vibhag Mishra Shala No. 8	22.19	0.134358453	
25	348	Sayajiganj Mishra Shala No. 48	28.23	0.140845208	
26	338	Sayajiganj Kanya Shala No. 38	29.22	0.144582844	
27	617	Wadi Marathi Mishra Shala No. 17	35.67	0.172061028	
28	213	Fatehpura Kumar Shala No. 13	18.09	0.17400346	
29	314	Sayajiganj Kumar Shala No. 14	22.93	0.182726341	
30	623	Wadi Kanya Shala No. 23	28.91	0.186591564	
31	624	Wadi Mishra Shala No. 24	21.52	0.208067391	
32	344	Sayajiganj Mishra Shala No. 44	25.91	0.209175575	
33	633	Wadi Kanya Shala No. 33	29.51	0.227463469	

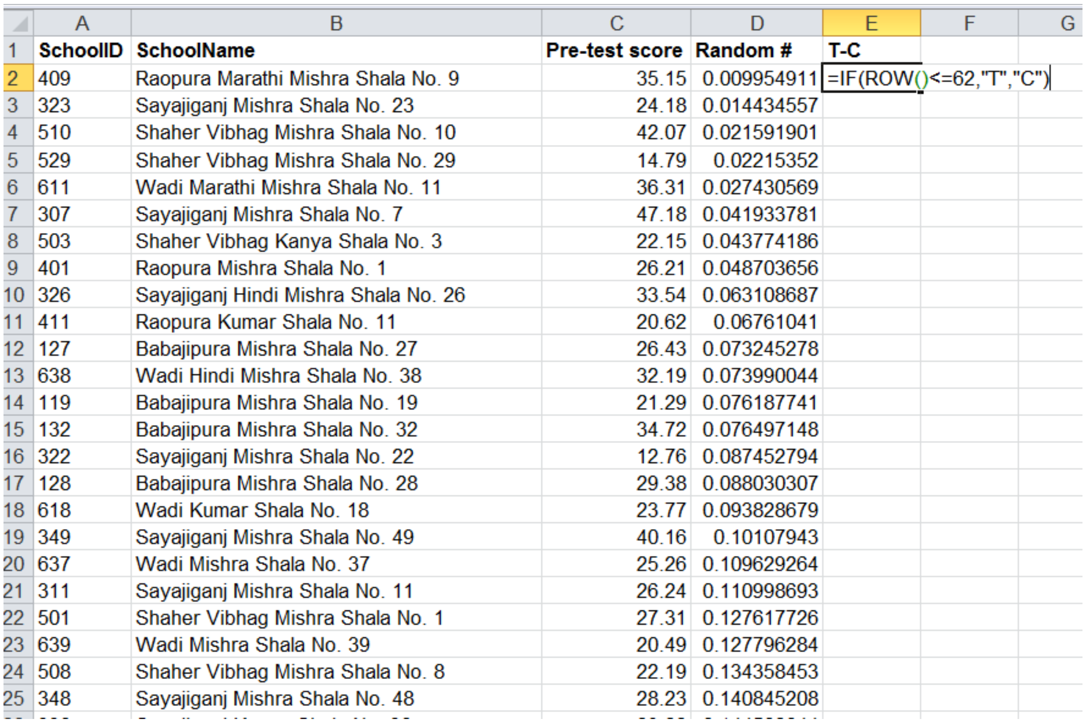
Because your list is randomly sorted, it is completely random whether schools are in the top half of the list, or the bottom half. Therefore, if you assign the top half to the treatment group and the bottom half to the control group, your schools have been "randomly assigned."

Step 4: Assign treatment/control status for each group

There are two ways to do this. To do this manually, in column E, type "T" for the first half of the rows (rows 2–63) and for the second half of the rows (rows 62–123), type "C". You can also do this by using the IF and MEDIAN functions. In Column E type:

```
=IF(D2<=MEDIAN($D$2:$D$123),"T","C")
```

And click and drag (or copy and paste) to the entire column. This will enter a "T" for schools that are below or at the median of the random number and a "C" for schools that are above it.



	A	B	C	D	E	F	G
1	SchoolID	SchoolName	Pre-test score	Random #	T-C		
2	409	Raopura Marathi Mishra Shala No. 9	35.15	0.009954911	=IF(ROW()<=62,"T","C")		
3	323	Sayajiganj Mishra Shala No. 23	24.18	0.014434557			
4	510	Shaher Vibhag Mishra Shala No. 10	42.07	0.021591901			
5	529	Shaher Vibhag Mishra Shala No. 29	14.79	0.02215352			
6	611	Wadi Marathi Mishra Shala No. 11	36.31	0.027430569			
7	307	Sayajiganj Mishra Shala No. 7	47.18	0.041933781			
8	503	Shaher Vibhag Kanya Shala No. 3	22.15	0.043774186			
9	401	Raopura Mishra Shala No. 1	26.21	0.048703656			
10	326	Sayajiganj Hindi Mishra Shala No. 26	33.54	0.063108687			
11	411	Raopura Kumar Shala No. 11	20.62	0.06761041			
12	127	Babajipura Mishra Shala No. 27	26.43	0.073245278			
13	638	Wadi Hindi Mishra Shala No. 38	32.19	0.073990044			
14	119	Babajipura Mishra Shala No. 19	21.29	0.076187741			
15	132	Babajipura Mishra Shala No. 32	34.72	0.076497148			
16	322	Sayajiganj Mishra Shala No. 22	12.76	0.087452794			
17	128	Babajipura Mishra Shala No. 28	29.38	0.088030307			
18	618	Wadi Kumar Shala No. 18	23.77	0.093828679			
19	349	Sayajiganj Mishra Shala No. 49	40.16	0.10107943			
20	637	Wadi Mishra Shala No. 37	25.26	0.109629264			
21	311	Sayajiganj Mishra Shala No. 11	26.24	0.110998693			
22	501	Shaher Vibhag Mishra Shala No. 1	27.31	0.127617726			
23	639	Wadi Mishra Shala No. 39	20.49	0.127796284			
24	508	Shaher Vibhag Mishra Shala No. 8	22.19	0.134358453			
25	348	Sayajiganj Mishra Shala No. 48	28.23	0.140845208			

Now select columns A through E and re-sort your list back in order of "SchoolID." You'll see that your schools have been randomly assigned to treatment and control groups.

Is the number of schools in both groups the same? Does the average pre-test score look balanced between the two groups?

Note that the number of schools in treatment and control will remain the same each time you re-randomize. This is because you are making sure that you always assign half of them to treatment and half to control. To check this, repeat step 1, but this time instead of copy pasting values, press the F9 key to re-randomize. Notice that the formula in column E will automatically recalculate the median each time and re-assign treatment and control status. Re-randomize 10 times and see what happens to the number of schools and the average pre-test score in each group.

Does the number of schools change when you re-randomize? Does the average pre-test score look balanced every time you re-randomize?

Part 3: stratified randomization

Stratification is the process of dividing a sample into groups, and then randomly assigning individuals within each group to the treatment and control. The reasons for doing this are rather technical. One reason for stratifying is that it ensures subgroups are balanced, making it easier to perform certain subgroup analyses. For example, if you want to test the effectiveness on a new education program separately for schools where children are taught in Hindi versus schools where children are taught in Gujarati, you can stratify by "language of instruction" and ensure that there are an equal number of schools of each language type in the treatment and control groups.

We have our list of schools and potential "strata"

Mechanically, the only difference in random sorting is that instead of simply sorting by the random number, you would first sort by language, and then the random number. Obviously, the first step is to ensure you have the variables by which you hope to stratify.

Step 1: Assign a random number to each school

Go to Column F and type:

=RAND()

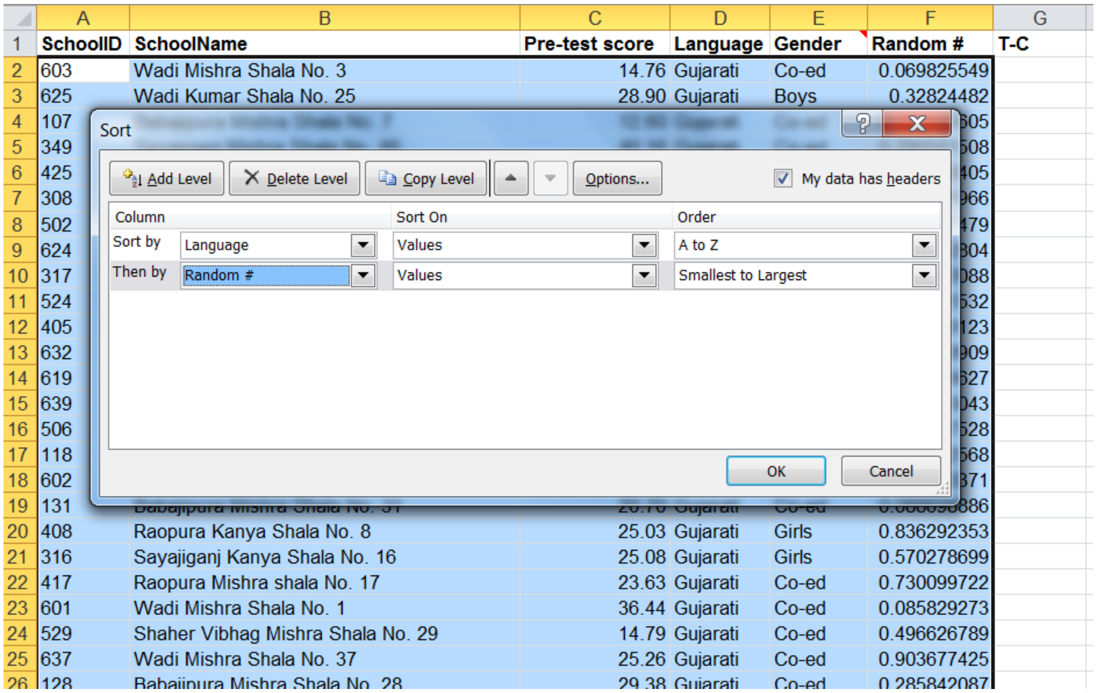
And click and drag (or copy and paste) to the entire column.

Step 2: Copy the cells in Column F, then paste the values over the same cells

Highlight all values in Column F. Then right-click anywhere in the highlighted column, and choose "Copy". Then, right-click anywhere in that column and choose "Paste Special."

Step 3: Sort by strata and then by random number

Assuming you have all the variables you need, you can now click "Sort" in the data tab. The Sort window will pop up. Sort by "Language." Press the "Add Level" button. Then select "Random #".



	A	B	C	D	E	F	G
1	SchoolID	SchoolName	Pre-test score	Language	Gender	Random #	T-C
2	603	Wadi Mishra Shala No. 3	14.76	Gujarati	Co-ed	0.069825549	
3	625	Wadi Kumar Shala No. 25	28.90	Gujarati	Boys	0.32824482	
4	107					605	
5	349					508	
6	425					405	
7	308					966	
8	502					479	
9	624					304	
10	317					088	
11	524					532	
12	405					123	
13	632					909	
14	619					627	
15	639					043	
16	506					528	
17	118					568	
18	602					371	
19	131	Babajipura Mishra Shala No. 31	20.70	Gujarati	Co-ed	0.066096886	
20	408	Raopura Kanya Shala No. 8	25.03	Gujarati	Girls	0.836292353	
21	316	Sayajiganj Kanya Shala No. 16	25.08	Gujarati	Girls	0.570278699	
22	417	Raopura Mishra shala No. 17	23.63	Gujarati	Co-ed	0.730099722	
23	601	Wadi Mishra Shala No. 1	36.44	Gujarati	Co-ed	0.085829273	
24	529	Shaher Vibhag Mishra Shala No. 29	14.79	Gujarati	Co-ed	0.496626789	
25	637	Wadi Mishra Shala No. 37	25.26	Gujarati	Co-ed	0.903677425	
26	128	Babajipura Mishra Shala No. 28	29.38	Gujarati	Co-ed	0.285842087	

Sort

☑ My data has headers

Column		Sort On	Order
Sort by	Language	Values	A to Z
Then by	Random #	Values	Smallest to Largest

Step 4: Assign treatment/control status for each group

There are two ways to do this. To do this manually, in column G, within each languages category, type "T" for the first half of the rows, and "C" for the second half. You can also do this by using the **IF** and **MEDIAN** functions. In Column G type:

```
=IF(F2<MEDIAN(IF($D$2:$D$123=D2,$F$2:$F$123)),"T","C")
```

And click and drag (or copy and paste) to the entire column. This will enter a "T" for schools that are below or at the median of the random number and a "C" for schools that are above it for *each* language category.

Is the total number of schools in both groups the same? Is the number of schools for each language category for both groups the same? Does the average pre-test score look balanced between the two groups?

Note that the total number of schools and the number of schools for each language category in treatment and control will remain the same each time you re-randomize. To check this, repeat step 1, but this time instead of copy pasting values, press the F9 key to re-randomize. Notice that the formula in column E will automatically recalculate the median for each category every time and re-assign treatment and control status. Re-randomize 10 times and see what happens to the number of schools and the average pre-test score in each group.

Does the total number of schools change when you re-randomize? Does the number of schools for each language category change? Does the average pre-test score look balanced every time you re-randomize?

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Exercise C: How to do Power Calculations in Optimal Design Software

CONTENTS

Key Vocabulary

1. POWER: The likelihood that, when a program/treatment has an effect, you will be able to distinguish the effect from zero i.e. from a situation where the program has no effect, given the sample size.

2. SIGNIFICANCE: The likelihood that the measured effect did not occur by chance. Statistical tests are performed to determine whether one group (e.g. the experimental group) is different from another group (e.g. comparison group) on certain outcome indicators of interest (for instance, test scores in an education program.)

3. STANDARD DEVIATION: For a particular indicator, a measure of the variation (or spread) of a sample or population. Mathematically, this is the square root of the variance.

4. STANDARDIZED EFFECT SIZE: A standardized (or normalized) measure of the [expected] magnitude of the effect of a program. Mathematically, it is the difference between the treatment and control group (or between any two treatment arms) for a particular outcome, divided by the standard deviation of that outcome in the control (or comparison) group.

5. **CLUSTER:** The unit of observation at which a sample size is randomized (e.g. school), each of which typically contains several units of observation that are measured (e.g. students). Generally, observations that are highly correlated with each other should be clustered and the estimated sample size required should be measured with an adjustment for clustering.

6. **INTRA-CLUSTER CORRELATION COEFFICIENT (ICC):** A measure of the correlation between observations within a cluster. For instance, if your experiment is clustered at the school level, the ICC would be the level of correlation in test scores for children in a given school relative to the overall correlation of students in all schools.

Introduction

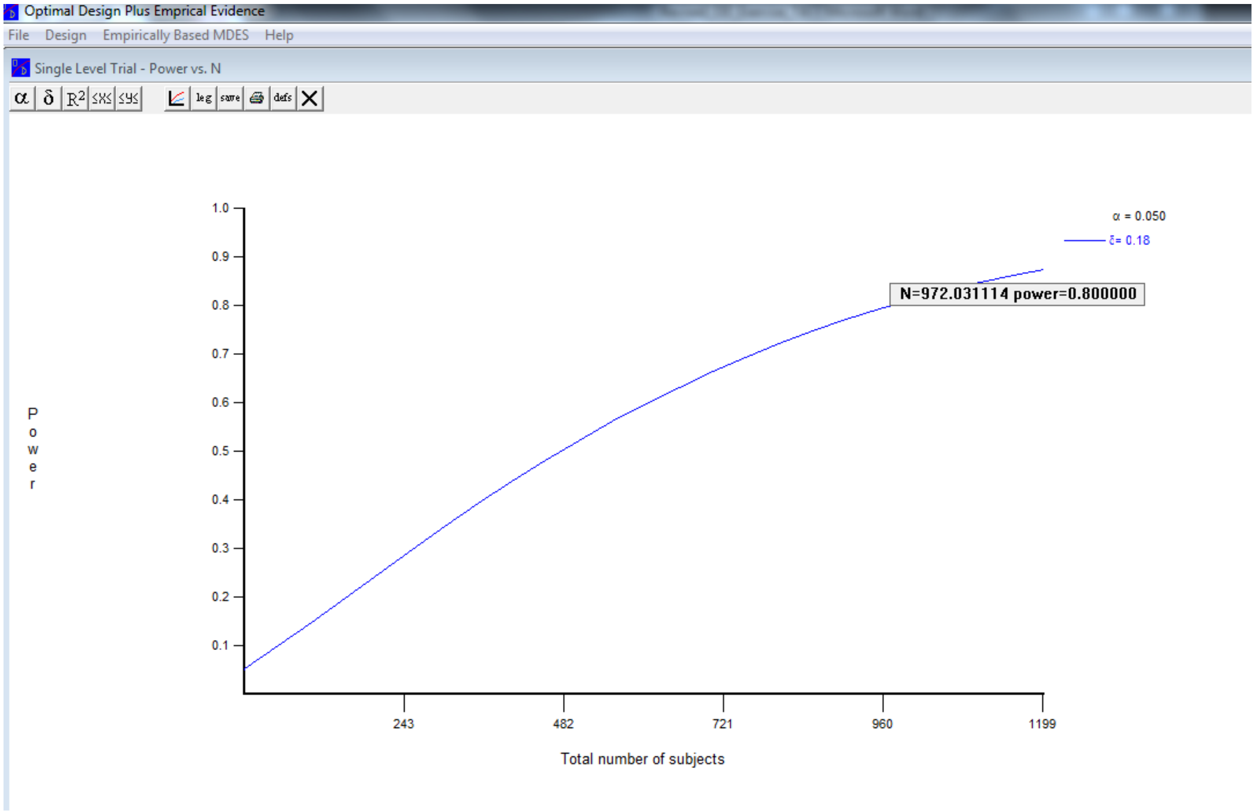
This exercise will help explain the trade-offs to power when designing a randomized trial. Should we sample every student in just a few schools? Should we sample a few students from many schools? How do we decide?

We will work through these questions by determining the sample size that allows us to detect a specific effect with at least 80 percent power, which is a commonly accepted level of power. Remember that power is the likelihood that when a program/treatment has an effect, you will be able to distinguish it from zero in your sample. Therefore at 80% power, if an intervention's impact is statistically significant at exactly the 5% level, then for a given sample size, we are 80% likely to detect an impact (i.e. we will be able to reject the null hypothesis.)

In going through this exercise, we will use the example of an education intervention that seeks to raise test scores. This exercise will demonstrate how the power of our sample changes with the number of school children, the number of children in each classroom, the expected magnitude of the change in test scores, and the extent to which children within a classroom behave more similarly than children across classrooms. We will use a software program called *Optimal* Design, developed by Stephen Raudenbush et al. with funding from the William T. Grant Foundation. Additional resources on research designs can be found on their web site.

Using the Optimal Design Software

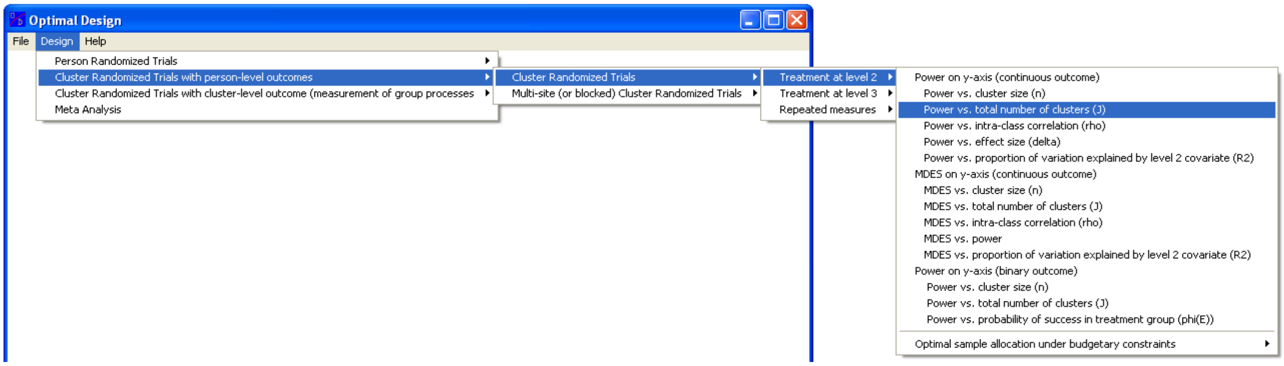
Optimal Design produces a graph that can show a number of comparisons: Power versus sample size (for a given effect), effect size versus sample size (for a given desired power), with many other options. The chart on the next page shows power on the y-axis and sample size on the x-axis. In this case, we inputted an effect size of 0.18 standard deviations (explained in the example that follows) and we see that we need a sample size of 972 to obtain a power of 80%.


Optimal Design Plus Emprical Evidence
File Design Empirically Based MDES Help
Single Level Trial - Power vs. N
α δ R² leg save defs
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
243
482
721
960
1199
Total number of subjects
α = 0.050
δ= 0.18
N=972.031114 power=0.800000

We will now go through a short example demonstrating how the OD software can be used to perform power calculations. If you haven't downloaded a copy of the OD software yet, you can do so from the following website (where a software manual is also available):

http://sitemaker.umich.edu/group-based/optimal_design_software

Running the HLM software file "od" should give you a screen which looks like the one below:


Optimal Design
File Design Help
Person Randomized Trials
Cluster Randomized Trials with person-level outcomes
Cluster Randomized Trials with cluster-level outcome (measurement of group processes
Meta Analysis
Cluster Randomized Trials
Multi-site (or blocked) Cluster Randomized Trials
Treatment at level 2
Treatment at level 3
Repeated measures
Power on y-axis (continuous outcome)
Power vs. cluster size (n)
Power vs. total number of clusters (J)
Power vs. intra-class correlation (rho)
Power vs. effect size (delta)
Power vs. proportion of variation explained by level 2 covariate (R2)
MDES on y-axis (continuous outcome)
MDES vs. cluster size (n)
MDES vs. total number of clusters (J)
MDES vs. intra-class correlation (rho)
MDES vs. power
MDES vs. proportion of variation explained by level 2 covariate (R2)
Power on y-axis (binary outcome)
Power vs. cluster size (n)
Power vs. total number of clusters (J)
Power vs. probability of success in treatment group (phi(E))
Optimal sample allocation under budgetary constraints

The various menu options under "Design" allow you to perform power calculations for randomized trials of various designs.

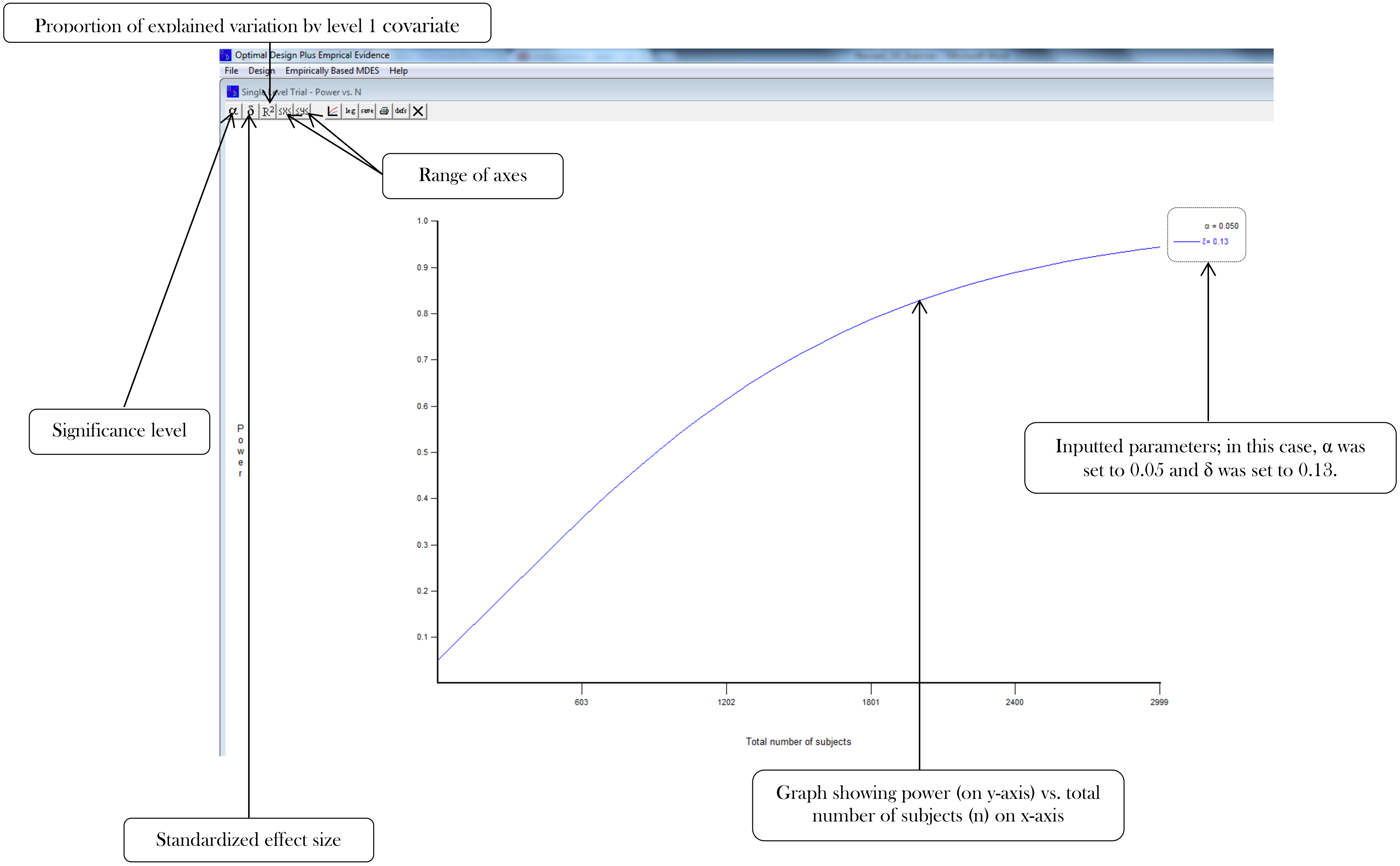
Let's work through an example that demonstrates how the sample size for a simple experiment can be calculated using OD. Follow the instructions along as you replicate the power calculations presented in this example, in OD. On the next page we have shown a sample OD graph, highlighting the various components that go into power calculations. These are:

- Significance level (**α**): For the significance level, typically denoted by α, the default value of 0.05 (i.e. a significance level of 95%) is commonly accepted.

- Standardized effect size (**δ**): Optimal Design (OD) requires that you input the standardized effect size, which is the effect size expressed in terms of a normal distribution with mean 0 and standard deviation 1. This will be explained in further detail below. The default value for δ is set to 0.200 in OD.

- Proportion of explained variation by level 1 covariate (**R^2**): This is the proportion of variation that you expect to be able to control for by including covariates (i.e. other explanatory variables other than the treatment) in your design or your specification. The default value for R^2 is set to 0 in OD.

- Range of axes (**≤x≤** and **≤y≤**): Changing the values here allows you to view a larger range in the resulting graph, which you will use to determine power.



Proportion of explained variation by level 1 covariate

Range of axes

Significance level

Inputted parameters; in this case, α was set to 0.05 and δ was set to 0.13.

Graph showing power (on y-axis) vs. total number of subjects (n) on x-axis

Standardized effect size

We will walk through each of these parameters below and the steps involved in doing a power calculation. Prior to that though, it is worth taking a step back to consider what one might call the “paradox of power”. Put simply, in order to perfectly calculate the sample size that your study will need, it is necessary to know a number of things: the effect of the program, the mean and standard deviation of your outcome indicator of interest for the control group, and a whole host of other factors that we deal with further on in the exercise. However, we cannot know or observe these final outcomes until we actually conduct the experiment! We are thus left with the following paradox: In order to conduct the experiment, we need to decide on a sample size…a decision that is contingent upon a number of outcomes that we cannot know without conducting the experiment in the first place.

It is in this regard that power calculations involve making careful assumptions about what the final outcomes are likely to be – for instance, what effect you realistically expect your program to have, or what you anticipate the average outcome for the control group being. These assumptions are often informed by real data: from previous studies of similar programs, pilot studies in your population of interest, etc. The main thing to note here is that to a certain extent, power calculations are more of an art than a science. However, making wrong assumptions will not affect accuracy (i.e, will not bias the results). It simply affects the precision with which you will be able to estimate your impact. Either way, it is useful to justify your assumptions, which requires carefully thinking through the details of your program and context.

With that said, let us work through the steps for a power calculation using an example. Say your research team is interested in looking at the impact of providing students a tutor. These tutors work with children in grades 2, 3 and 4 who are identified as falling behind their peers. Through a pilot survey, we know that the average test scores of students before receiving tutoring is 26 out of 100, with a standard deviation of 20. We are interested in evaluating whether tutoring can cause a 10 percent increase in test scores.

1) Let’s find out the minimum sample that you will need in order to be able to detect whether the tutoring program causes a 10 percent increase in test scores. Assume that you are randomizing at the school level i.e. there are treatment schools and control schools.

I. What will be the mean test score of members of the control group? What will the standard deviation be?

> **Answer:** To get the mean and standard deviation of the control group, we use the mean and standard deviation from our pilot survey i.e. mean = 26 and standard deviation = 20. Since we do not know how the control group’s scores will change, we assume that the control group’s scores will not increase absent the tutoring program and will correspond to the scores from our pilot data.

II. If the intervention is supposed to increase test scores by 10%, what should you expect the mean and standard deviation of the treatment group to be after the intervention? Remember, in this case we are considering a 10% increase in scores over the scores of the control group, which we calculated in part I.

> **Answer:** Given that the mean of the control group is 26, the mean with a 10% increase would be 26*1.10 = 28.6. With no information about the sample distribution of the treatment group after the

intervention, we have no reason for thinking that there is a higher amount of variability within the treatment group than the control group (i.e. we assume homogeneous treatment impacts across the population). In reality, the treatment is likely to have heterogeneous i.e. differential impacts across the population, yielding a different standard deviation for the treatment group. For now, we assume the standard deviation of the treatment group to be the same as that of the control group i.e. 20.

III. Optimal Design (OD) requires that you input the standardized effect size, which is the effect size expressed in terms of a normal distribution with mean 0 and standard deviation 1. Two of the most important ingredients in determining power are the *effect size* and the *variance* (or standard deviation). The standardized effect size basically combines these two ingredients into one number. The standardized effect size is typically denoted using the symbol δ (delta), and can be calculated using the following formula:

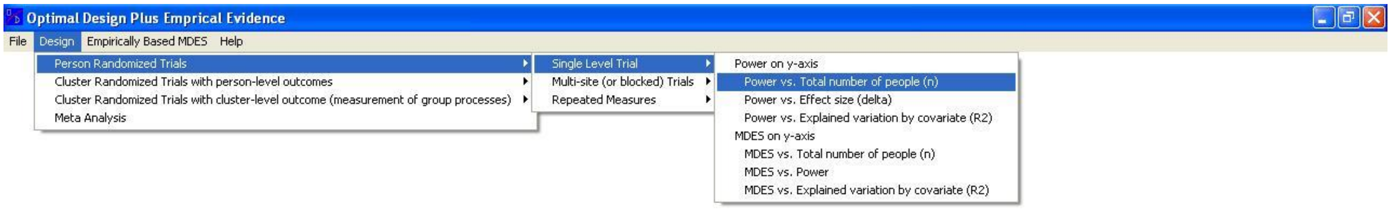
$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

Using this formula, what is δ?

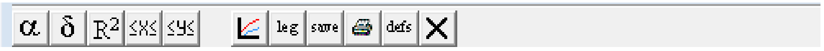
Answer: $\boldsymbol{\delta = \frac{(28.6-26)}{20} = 0.13}$

IV. Now use OD to calculate the sample size that you need in order to detect a 10% increase in test scores. You can do this by navigating in OD as follows:

Design → Person Randomized Trials → Single Level Trial → Power vs. Total number of people (n)


Optimal Design Plus Emprical Evidence
File
Design
Empirically Based MDES
Help
Person Randomized Trials
Cluster Randomized Trials with person-level outcomes
Cluster Randomized Trials with cluster-level outcome (measurement of group processes)
Meta Analysis
Single Level Trial
Multi-site (or blocked) Trials
Repeated Measures
Power on y-axis
Power vs. Total number of people (n)
Power vs. Effect size (delta)
Power vs. Explained variation by covariate (R2)
MDES on y-axis
MDES vs. Total number of people (n)
MDES vs. Power
MDES vs. Explained variation by covariate (R2)

There are various parameters that you will be asked to fill in:



You can do this by clicking on the button with the symbol of the parameter. To reiterate, the parameters are:

- Significance level (**α**): For the significance level, typically denoted by α, the default value of 0.05 (i.e. a significance level of 95%) is commonly accepted.

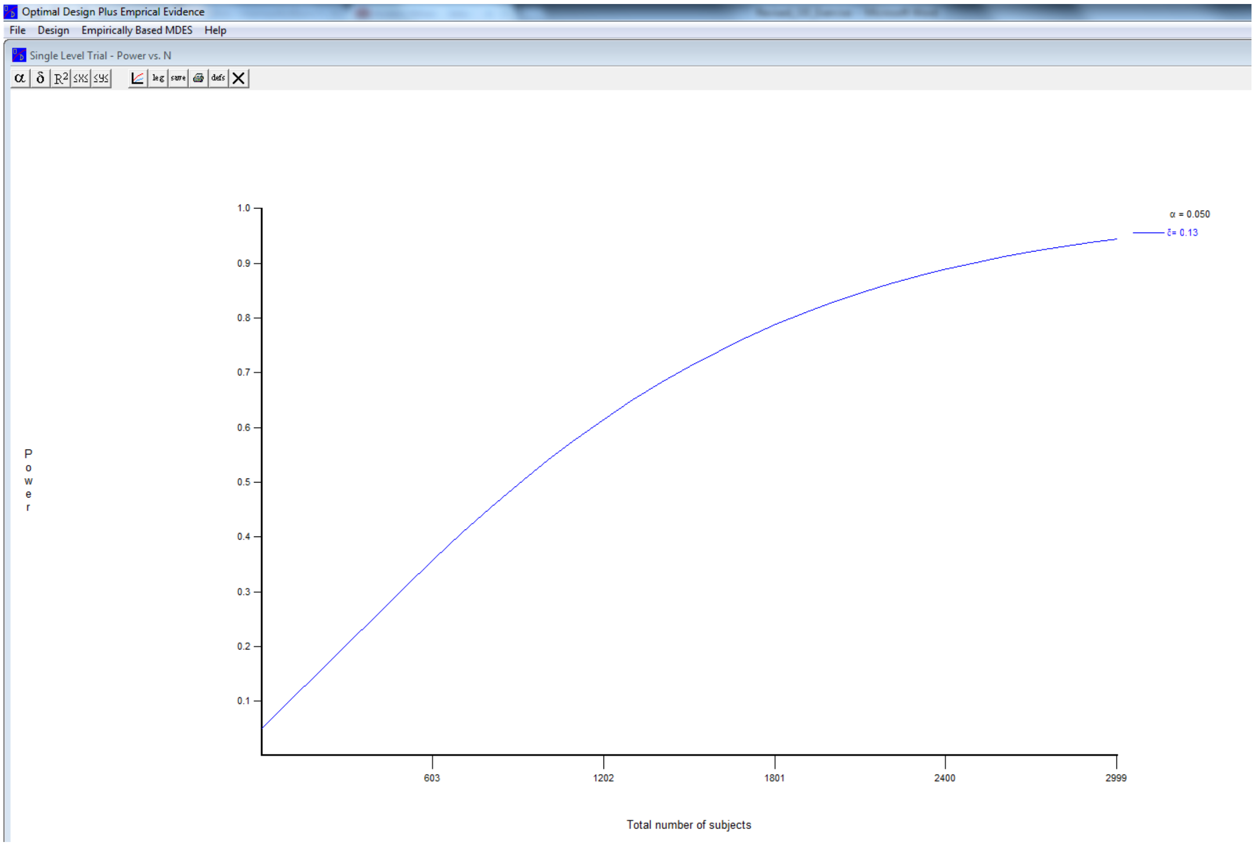
- Standardized effect size (**δ**): The default value for δ is set to 0.200 in OD. However, you will want to change this to the value that we computed for δ in part C.

- Proportion of explained variation by level 1 covariate (**R^2**): This is the proportion of variation that you expect to be able to control for by including covariates (i.e. other explanatory variables other than the treatment) in your design or your specification. We will leave this at the default value of 0 for now and return to it later on.

- Range of axes (**≤x≤** and **≤y≤**): Changing the values here allows you to view a larger range in the resulting graph, which you will use to determine power; we will return to this later, but can leave them at the default values for now.

What will your total sample size need to be in order to detect a 10% increase in test scores at 80% power?

> **Answer:** Once you input the various values above into the appropriate cells, you will get a plot with power on the y-axis and the total number of subjects on the x-axis. Click your mouse on the plot to see the power and sample size for any given point on the line.
>
> Power of 80% (0.80 on the y-axis of your chart) is typically considered an acceptable threshold. This is the level of power that you should aim for while performing your power calculations. You will notice that just inputting the various values above does not allow you to see the number of subjects required for 80% power. You will thus need to increase the range of your x-axis; set the maximum value at 3000. This will yield a plot that looks the one on the following page.
>
> To determine the sample size for a given level of power, click your mouse cursor on the graph line at the appropriate point. While this means that arriving at *exactly* a given level of power (say power of exactly 0.80) is difficult, a very good approximate (i.e. within a couple of decimal places) is sufficient for our purposes.
>
> Clicking your mouse cursor on the line at the point where Power ~ 0.8 tells us that the total number of subjects, called "N", is approximately 1,850. OD assumes that the sample will be balanced between the treatment and control groups. **Thus, the treatment group will have 1850/2 = 925 students and the control group will have 1850/2 = 925 students as well.**



Estimating Sample Size for a Simple Experiment

All right, now it is your turn! For the parts A – I below, leave the value of R^2 at the default of 0 whenever you use OD; we will experiment with changes in the R^2 value a little later.

You decide that you would like your study to be powered to measure an increase in test scores of 20% rather than 10%. Try going through the steps that we went through in the example above. Let's find out the minimum sample you will need in order to detect whether the tutoring program can increase test scores by 20%.

A. What is the mean test score for the control group? What is the standard deviation?

Mean:

Standard deviation:

B. If the intervention is supposed to increase test scores by 20%, what should you expect the mean and standard deviation of the treatment group to be after the intervention?

Mean:

Standard deviation:

C. What is the desired standardized effect size δ? Remember, the formula for calculating δ is:

$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

δ:

D. Now use OD to calculate the sample size that you need in order to detect a 20% increase in test scores.

Sample size (n):

Treatment:

Control:

E. Is the *minimum* sample size required to detect a 10% increase in test scores larger or smaller than the minimum sample size required to detect a 20% increase in test scores? Intuitively, will you need larger or smaller samples to measure smaller effect sizes?

Answer:

F. Your research team has been thrown into a state of confusion! While one prior study led you to believe that a 20% increase in test scores is possible, a recently published study suggests that a more conservative 10% increase is more plausible. What sample size should you pick for your study?

Answer:

G. Both the studies mentioned in part F found that although average test scores increased after the tutoring intervention, the standard deviation of test scores also increased i.e. there was a larger spread of test scores across the treatment groups. To account for this, you posit that instead of 20, the standard deviation of test scores may now be 25 after the tutoring program. Calculate the new δ for an increase of 10% in test scores.

δ:

H. For an effect of 10% on test scores, does the corresponding standardized effect size increase, decrease, or remain the same if the standard deviation is 25 versus 20? Without plugging the values into OD, all other things being equal, what impact does a higher standard deviation of your outcome of interest have on your required sample size?

Answer:

I. Having gone through the intuition, now use OD to calculate the sample size required in order to detect a 10% increase in test scores, if the pre-intervention mean test scores are 26, with a standard deviation of 25.

Sample size (n):

Treatment:

Control:

J. One way by which you can increase your power is to include covariates i.e. control variables that you expect will explain some part of the variation in your outcome of interest. For instance, baseline, pre-intervention test scores may be a strong predictor of a child's post-intervention test scores; including baseline test scores in your eventual regression specification would help you to isolate the variation in test scores attributable to the tutoring intervention more precisely. You can account for the presence of covariates in your power calculations using the R^2 parameter, in which you specify what proportion of the eventual variation in your outcome of interest is attributable to your treatment condition.

Say that you have access to the pre-intervention test scores of children in your sample for the tutoring study. Moreover, you expect that pre-intervention test scores explain 50% of the variation in post-intervention scores. What size sample will you require in order to measure an increase in test scores of 10%, assuming standard deviation in test scores of 25, with a pre-intervention mean of 26. Is this more or less than the sample size that you calculated in part I?

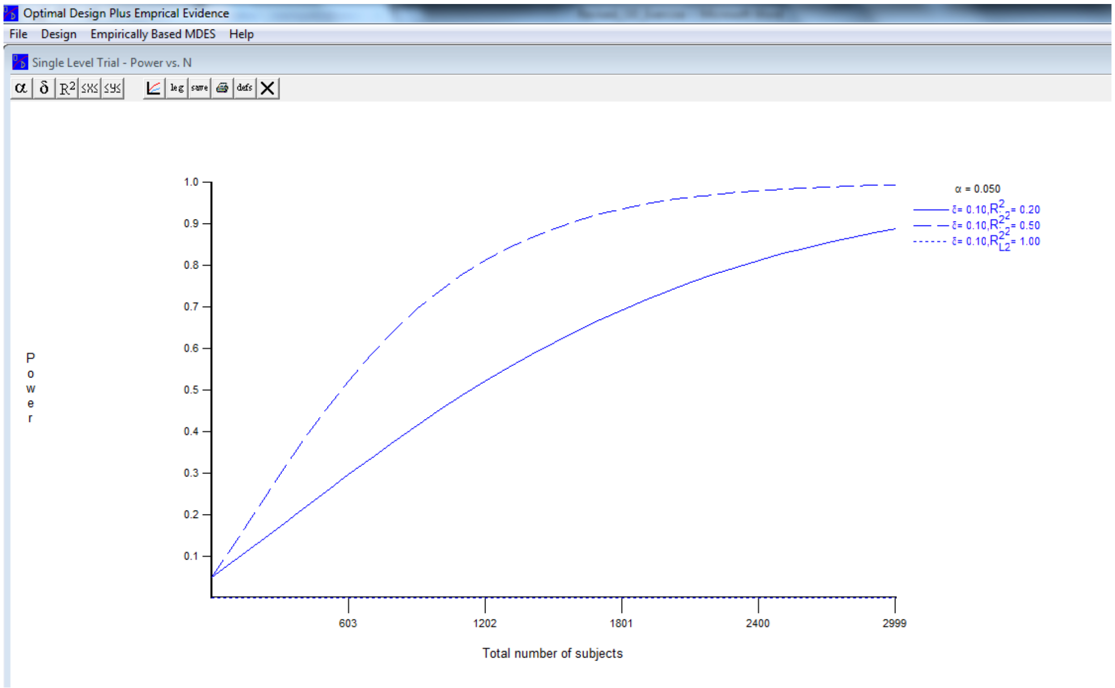
Sample size (n):

Treatment:

Control:

K. One of your colleagues on the research team thinks that 50% may be too ambitious an estimate of how much of the variation in test scores post-intervention is attributable to baseline scores. She suggests that 20% may be a better estimate. What happens to your required sample size when you run the calculations from part J with an R^2 of 0.200 instead of 0.500? What happens if you set R^2 to be 1.000?

Tip: You can enter up to 3 separate values on the same graph for the R^2 in OD; if you do, you will end up with a figure like the one below:



Answer:

Some Wrinkles: Limited Resources and Imperfect Compliance

L. You find out that you only have enough funds to survey 1,200 children. Assume that you do not have data on baseline covariates, but know that pre-intervention test scores were 26 on average, with a standard deviation of 20. What standardized effect size (δ) would you need to observe in order to survey a maximum of 1,200 children and still retain 80% power? Assume that the R^2 is 0 for this exercise since you have no baseline covariate data.

Hint: You will need to plot "Power vs. Effect size (delta)" in OD, setting "N" to 1,200. You can do this by navigating in OD as follows: Design → Person Randomized Trials → Single Level Trial → Power *vs. Effect Size (delta). Then, click on the point of your graph that roughly corresponds to power = 0.80 on the y-axis.*

δ =

M. Your research team estimates that you will not realistically see more than a 10% increase in test scores due to the intervention. Given this information, is it worth carrying out the study on just 1,200 children if you are adamant about still being powered at 80%?

Answer:

N. Your research team is hit with a crisis: You are told that you cannot force people to use the tutors! After some small focus groups, you estimate that only 40% of schoolchildren would be interested in the tutoring services. You realize that this intervention would only work for a very limited number of schoolchildren. You do not know in advance whether students are likely to take up the tutoring service or not. How does this affect your power calculations?

Answer:

O. You have to "adjust" the effect size you want to detect by the proportion of individuals that actually gets treated. Based on this, what will be your "adjusted" effect size and the adjusted standardized effect size (δ) if you originally wanted to measure a 10% increase in test scores? Assume that your pre-intervention mean test score is 26, with a standard deviation of 20, you do not have any data on covariates, and that you can survey as many children as you want.

Hint: Keep in mind that we are calculating the average treatment effect for the entire group here. Thus, the lower the number of children that actually receives the tutoring intervention, the lower will be the measured effect size.

Answer:

P. What sample size will you need in order to measure the effect size that you calculated in part O with 80% power? Is this sample bigger or smaller than the sample required when you assume that 100% of children take up the tutoring intervention (as we did in the example at the start)?

Sample size (n):

Treatment:

Control:

Clustered Designs

Thus far we have considered a simple design where we randomize at the *individual-level* i.e. school children are either assigned to the treatment (tutoring) or control (no tutoring) condition. However, spillovers could be a major concern with such a design: If treatment and control students are in the same *school*, let alone the same classroom, students receiving tutoring may affect the outcomes for students not receiving tutoring (through peer learning effects) and vice versa. This would lead us to get a biased estimate of the impact of the tutoring program.

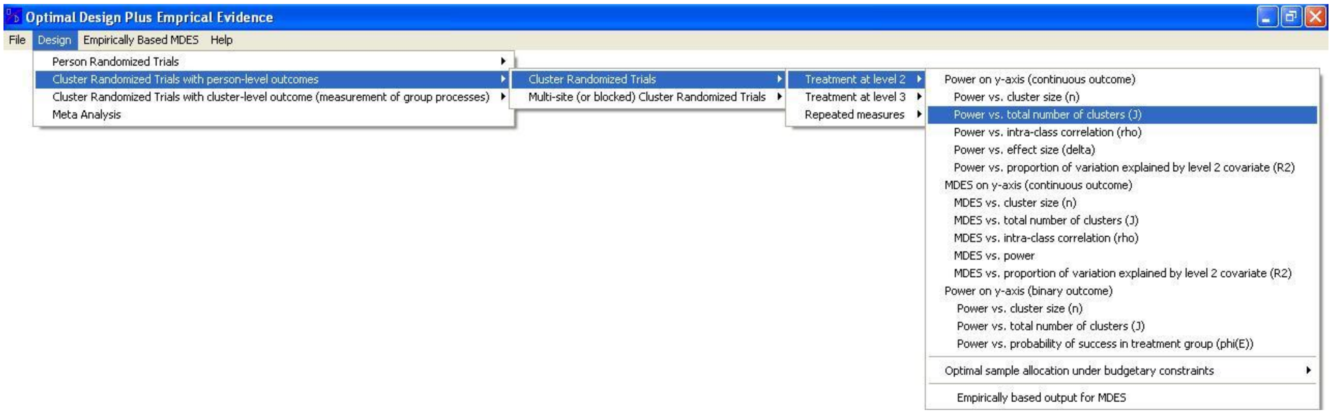
In order to preclude this, your research team decides that it would like to run a cluster randomized trial, randomizing at the *school-level* instead of the individual-level. In this case, each school forms a "cluster", with all the students in a given school assigned to either the treatment condition, or the control one. Under such a design, the only spillovers that may show up would be across schools, a far less likely possibility than spillovers within schools.

Since the behavior of individuals in a given cluster will be correlated, we need to take an **intra-cluster or intra-class correlation (denoted by the Greek symbol ρ)** into account for each outcome variable of interest. Remember, ρ is a measure of the correlation between children within a given school (see key vocabulary at the start of this exercise.) ρ tells us how strongly the outcomes are correlated for units within the same cluster. If students from the same school were clones (no variation) and all scored the same on the test, then ρ would equal 1. If, on the other hand, students from the same schools were in fact independent and there was zero difference between schools or any other factor that affected those students, then ρ would equal 0.

The ρ or ICC of a given variable is typically determined by looking at pilot or baseline data for your population of interest. Should you not have the data, another way of estimating the ρ is to look at other studies examining similar outcomes amongst similar populations. Given the inherent uncertainty with this, it is useful to consider a range of ρs when conducting your power calculations (a sensitivity analysis) to see how sensitive they are to changes in ρ. We will look at this a little further on. While the ρ can vary widely depending on what you are looking at, values of less than 0.05 are typically considered low, values between 0.05-0.20 are considered to be of moderate size, and values above 0.20 are considered fairly high. Again, what counts as a low ρ and what counts as a high ρ can vary dramatically by context and outcome of interest, but these ranges can serve as initial rules of thumb.

Based on a pilot study and earlier tutoring interventions, your research team has determined that the ρ is 0.17. You need to calculate the total sample size to measure a 15% increase in test scores (assuming that test scores at the baseline are 26 on average, with a standard deviation of 20, setting R^2 to 0 for now). You can do this by navigating in OD as follows:

Design → Cluster Randomized Trials with person-level outcomes → Cluster Randomized Trials → Treatment at Level 2 → Power vs. total number of clusters (J)


Optimal Design Plus Emprical Evidence
File
Design
Empirically Based MDES
Help
Person Randomized Trials
Cluster Randomized Trials with person-level outcomes
Cluster Randomized Trials with cluster-level outcome (measurement of group processes)
Meta Analysis
Cluster Randomized Trials
Multi-site (or blocked) Cluster Randomized Trials
Treatment at level 2
Treatment at level 3
Repeated measures
Power on y-axis (continuous outcome)
Power vs. cluster size (n)
Power vs. total number of clusters (J)
Power vs. intra-class correlation (rho)
Power vs. effect size (delta)
Power vs. proportion of variation explained by level 2 covariate (R2)
MDES on y-axis (continuous outcome)
MDES vs. cluster size (n)
MDES vs. total number of clusters (J)
MDES vs. intra-class correlation (rho)
MDES vs. power
MDES vs. proportion of variation explained by level 2 covariate (R2)
Power on y-axis (binary outcome)
Power vs. cluster size (n)
Power vs. total number of clusters (J)
Power vs. probability of success in treatment group (phi(E))
Optimal sample allocation under budgetary constraints
Empirically based output for MDES

In the bar at the top, you will see the same parameters as before, with an additional option for the intra-cluster correlation. Note that OD uses "n" to denote the cluster size here, not the total sample size. OD assigns two default values for the effect size (δ) and the intra-cluster correlation (ρ), so do not be alarmed if you see four lines on the chart. Simply delete the default values and replace them with the values for the effect size and intra-cluster correlation that you are using.

Q. What is the effect size (δ) that you want to detect here? Remember that the formula for calculating δ is:

$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

δ:

R. Assuming there are 40 children per school, how many schools would you need in your clustered randomized trial?

Answer:

S. Given your answer above, what will the total size of your sample be?

Sample size:

Treatment:

Control:

T. What would the number of schools and total sample size be if you assumed that 20 children from each school were part of the sample? What about if 100 children from each school were part of the sample?

	20 children per school	40 children per school	100 children per school
Number of schools:		160	
Total no. of students:		6,400	

U. As the number of clusters increases, does the total number of students required for your study increase or decrease? Why do you suspect this is the case? What happens as the number of children per school increases?

Answer:

V. You realize that you had read the pilot data wrong: It turns out that the ρ is actually 0.07 and not 0.17. Now what would the number of schools and total sample size be if you assumed that 20 children from each school were part of the sample? What about if 40 or 100 children from each school were part of the sample?

	20 children per school	40 children per school	100 children per school
Number of schools:			
Total no. of students:			

W. How does the total sample size change as you increase the number of individuals per cluster in part V? How do your answers here compare to your answers in part T?

Answer:

X. Given a choice between offering the tutors to more children in each school (i.e. adding more individuals to the cluster) versus offering tutors in more schools (i.e. adding more clusters), which option is best *purely from the perspective of improving statistical power?* Can you imagine a situation when there will not be much difference between the two from the perspective of power?

Answer:

Group Presentation

Participants will form 4-6 person groups which will work through the design process for a randomized evaluation of a development project. Groups will be aided in this project by both the faculty and teaching assistants with the work culminating in presentations at the end of the week.

The goal of the group presentations is to consolidate and apply the knowledge of the lectures and thereby ensure that participants will leave with the knowledge, experience, and confidence necessary to conduct their own randomized evaluations. We encourage groups to work on projects that are relevant to participants' organisations.

All groups will present on Friday. The 15-minute presentation is followed by a 15-minute discussion led by J-PAL affiliates and staff. We provide groups with template slides for their presentation (see next page). While the groups do not need to follow this exactly, the presentation should have no more than 9 slides (including title slide, excluding appendix) and should include the following topics:

- Brief project background
- Theory of change
- Evaluation question
- Outcomes
- Evaluation design
- Data and sample size
- Potential validity threats and how to manage them
- Dissemination strategy of results

Please time yourself and do not exceed the allotted time. We have only a limited amount of time for these presentations, so we will follow a strict timeline to be fair to all groups.

Title

List your Team Members

You don’t have to follow this
exactly, this is just a guideline.

Background

- Talk briefly about general context, needs assessment, problem you want to solve.

Theory of Change

- Describe the specific intervention that you are evaluating.
- Talk about how it will solve part of the problem you described in the background.
- You may want to mention other causes of a problem that your intervention will not solve.
- (You can use the TOC template in the appendix.)

Evaluation Questions and Outcomes

- These should be directly linked to the TOC described above.
- What outcomes do you need to measure to test your research hypothesis?

Evaluation Design

- Unit of randomization, type of randomization (why did you choose these?)
- The actual randomization design- i.e. specific treatment group(s)

Data and Sample Size

- Outcomes
- Tell us where you will get the data – survey? Administrative?
- Power calcs
  - Justify where you got effect size and rho from, don’t make it up.
  - You may need to do separate power calcs for separate outcomes.

Potential challenges

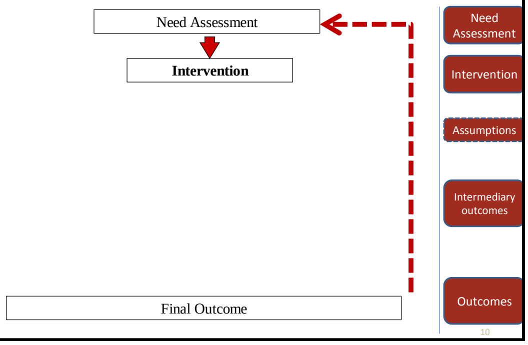
- Talk about threats (attrition, spillover, etc.) and how you want to manage them.
- You may need to revise your power calcs.

Results

- Why (and for whom) they would be useful.
- How would you disseminate them.

Appendix

Theory of change



Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence

Coalition for Evidence-Based Policy

A NONPROFIT, NONPARTISAN ORGANIZATION

Updated February 2010

This publication was produced by the Coalition for Evidence-Based Policy, with funding support from the William T. Grant Foundation, Edna McConnell Clark Foundation, and Jerry Lee Foundation.

We welcome comments and suggestions on this document (jbaron@coalition4evidence.org).

Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence

This is a checklist of key items to look for in reading the results of a randomized controlled trial of a social program, project, or strategy (“intervention”), to assess whether it produced valid evidence on the intervention’s effectiveness. This checklist closely tracks guidance from both the U.S. Office of Management and Budget (OMB) and the U.S. Education Department’s Institute of Education Sciences (IES)[1]; however, the views expressed herein do not necessarily reflect the views of OMB or IES.

This checklist limits itself to key items, and does not try to address all contingencies that may affect the validity of a study’s results. It is meant to aid – not substitute for – good judgment, which may be needed for example to gauge whether a deviation from one or more checklist items is serious enough to undermine the study’s findings.

A brief appendix addresses *how many* well-conducted randomized controlled trials are needed to produce strong evidence that an intervention is effective.

Checklist for overall study design

□ **Random assignment was conducted at the appropriate level – either groups (e.g., classrooms, housing projects), or individuals (e.g., students, housing tenants), or both.**

Random assignment of individuals is usually the most efficient and least expensive approach. However, it may be necessary to randomly assign groups – instead of, or in addition to, individuals – in order to evaluate (i) interventions that may have sizeable “spillover” effects on nonparticipants, and (ii) interventions that are delivered to whole groups such as classrooms, housing projects, or communities. (See reference 2 for additional detail.[2])

□ **The study had an adequate sample size – one large enough to detect meaningful effects of the intervention.**

Whether the sample is sufficiently large depends on specific features of the intervention, the sample population, and the study design, as discussed elsewhere.[3] Here are two items that can help you judge whether the study you’re reading had an adequate sample size:

- If the study found that the intervention produced *statistically-significant* effects (as discussed later in this checklist), then you can probably assume that the sample was large enough.
- If the study found that the intervention did *not* produce statistically-significant effects, the study report should include an analysis showing that the sample was large enough to detect meaningful effects of the intervention. (Such an analysis is known as a “power” analysis.[4])

Reference 5 contains illustrative examples of sample sizes from well-conducted randomized controlled trials conducted in various areas of social policy.[5]

Checklist to ensure that the intervention and control groups remained equivalent during the study

□ **The study report shows that the intervention and control groups were highly similar in key characteristics prior to the intervention (e.g., demographics, behavior).**

□ **If the study asked sample members to consent to study participation, they provided such consent *before* learning whether they were assigned to the intervention versus control group.**

If they provided consent afterward, their knowledge of which group they are in could have affected their decision on whether to consent, thus undermining the equivalence of the two groups.

□ **Few or no control group members participated in the intervention, or otherwise benefited from it (i.e., there was minimal "cross-over" or "contamination" of controls).**

□ **The study collected outcome data in the same way, and at the same time, from intervention and control group members.**

□ **The study obtained outcome data for a high proportion of the sample members originally randomized (i.e., the study had low sample "attrition").**

As a general guideline, the studies should obtain outcome data for at least 80 percent of the sample members originally randomized, including members assigned to the intervention group who did not participate in or complete the intervention. Furthermore, the follow-up rate should be approximately the same for the intervention and the control groups.

The study report should include an analysis showing that sample attrition (if any) did not undermine the equivalence of the intervention and control groups.

□ **The study, in estimating the effects of the intervention, kept sample members in the original group to which they were randomly assigned.** This even applies to:

- Intervention group members who failed to participate in or complete the intervention (retaining them in the intervention group is consistent with an "intention-to-treat" approach); and

- Control group members who may have participated in or benefited from the intervention (i.e., "cross-overs," or "contaminated" members of the control group).[6]

Checklist for the study's outcome measures

□ **The study used "valid" outcome measures – i.e., outcome measures that are highly correlated with the true outcomes that the intervention seeks to affect.** For example:

- Tests that the study used to measure outcomes (e.g., tests of academic achievement or psychological well-being) are ones whose ability to measure true outcomes is well-established.

- If sample members were asked to self-report outcomes (e.g., criminal behavior), their reports were corroborated with independent and/or objective measures if possible (e.g., police records).

- The outcome measures did not favor the intervention group over the control group, or vice-versa. For instance, a study of a computerized program to teach mathematics to young students should not measure outcomes using a computerized test, since the intervention group will likely have greater facility with the computer than the control group.[7]

□ **The study measured outcomes that are of policy or practical importance – not just intermediate outcomes that may or may not predict important outcomes.**

As illustrative examples: (i) the study of a pregnancy prevention program should measure outcomes such as actual pregnancies, and not just participants' attitudes toward sex; and (ii) the study of a remedial reading program should measure outcomes such as reading comprehension, and not just the ability to sound out words.

□ **Where appropriate, the members of the study team who collected outcome data were "blinded" – i.e., kept unaware of who was in the intervention and control groups.**

Blinding is important when the study measures outcomes using interviews, tests, or other instruments that are not fully structured, possibly allowing the person doing the measuring some room for subjective judgment. Blinding protects against the possibility that the measurer's bias (e.g., as a proponent of the intervention) might influence his or her outcome measurements. Blinding would be important, for example, in a study that measures the incidence of hitting on the playground through playground observations, or a study that measures the word identification skills of first graders through individually-administered tests.

□ **Preferably, the study measured whether the intervention's effects lasted long enough to constitute meaningful improvement in participants' lives (e.g., a year, hopefully longer).**

This is important because initial intervention effects often diminish over time – for example, as changes in intervention group behavior wane, or as the control group "catches up" on their own.

Checklist for the study's reporting of the intervention's effects

□ **If the study claims that the intervention has an effect on outcomes, it reports (i) the size of the effect, and whether the size is of policy or practical importance; and (ii) tests showing the effect is statistically significant (i.e., unlikely to be due to chance).**

These tests for statistical significance should take into account key features of the study design, including:

- Whether individuals (e.g., students) or groups (e.g., classrooms) were randomly assigned;

- Whether the sample was sorted into groups prior to randomization (i.e., "stratified," "blocked," or "paired"); and

- Whether the study intends its estimates of the intervention's effect to apply only to the sites (e.g., housing projects) in the study, or to be generalizable to a larger population.

- □ **The study reports the intervention's effects on all the outcomes that the study measured, not just those for which there is a positive effect.**

  This is so you can gauge whether any positive effects are the exception or the pattern. In addition, if the study found only a limited number of statistically-significant effects among many outcomes measured, it should report tests showing that such effects were unlikely to have occurred by chance.

Appendix: How many randomized controlled trials are needed to produce strong evidence of effectiveness?

To have strong confidence that an intervention would be effective if faithfully replicated, one generally would look for evidence including the following:

- □ **The intervention has been demonstrated effective, through well-conducted randomized controlled trials, in more than one site of implementation.**

  Such a demonstration might consist of two or more trials conducted in different implementation sites, or alternatively one large multi-site trial.

- □ **The trial(s) evaluated the intervention in the real-world community settings and conditions where it would normally be implemented** (e.g., community drug abuse clinics, public schools, job training program sites).

  This is as opposed to tightly-controlled conditions, such as specialized sites that researchers set up at a university for purposes of the study, or settings where the researchers themselves administer the intervention.

- □ **There is no strong countervailing evidence, such as well-conducted randomized controlled trials of the intervention showing an absence of effects.**

References

[1] U.S. Office of Management and Budget (OMB), What Constitutes Strong Evidence of Program Effectiveness, http://www.whitehouse.gov/omb/part/2004_program_eval.pdf, 2004; U.S. Department of Education's Institute of Education Sciences, Identifying and Implementing Educational Practices Supported By Rigorous Evidence, http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html, December 2003; What Works Clearinghouse of the U.S. Education Department's Institute of Education Sciences, Key Items To Get Right When Conducting A Randomized Controlled Trial in Education, prepared by the Coalition for Evidence-Based Policy, http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf.

[2] Random assignment of groups rather than, or in addition to, individuals may be necessary in situations such as the following:

(a) The intervention may have sizeable "spillover" effects on individuals other than those who receive it.

For example, if there is good reason to believe that a drug-abuse prevention program for youth in a public housing project may produce sizeable reductions in drug use not only among program participants, but also among their peers in the same housing project (through peer-influence), it is probably necessary to randomly assign whole housing projects to intervention and control groups to determine the program's effect. A study that only randomizes individual youth within a housing project to intervention versus control groups will underestimate the program's effect to the extent the program reduces drug use among both intervention and control-group students in the project.

(b) The intervention is delivered to groups such as classrooms or schools (e.g., a classroom curriculum or schoolwide reform program), and the study seeks to distinguish the effect of the intervention from the effect of other group characteristics (e.g., quality of the classroom teacher).

For example, in a study of a new classroom curriculum, classrooms in the sample will usually differ in two ways: (i) whether they use the new curriculum or not, and (ii) who is teaching the class. Therefore, if the study (for example) randomly assigns individual students to two classrooms that use the curriculum versus two classrooms that don't, the study will not be able to distinguish the effect of the curriculum from the effect of other classroom characteristics, such as the quality of the teacher. Such a study therefore probably needs to randomly assign whole classrooms and teachers (a sufficient sample of each) to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also in classroom and teacher characteristics.

For similar reasons, a study of a schoolwide reform program will probably need to randomly assign whole schools to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also school characteristics (e.g., teacher quality, average class size).

[3] What Works Clearinghouse of the U.S. Education Department's Institute of Education Sciences, *Key Items To Get Right When Conducting A Randomized Controlled Trial in Education*, op. cit., no. 1.

[4] Resources that may be helpful in reviewing or conducting power analyses include: the William T. Grant Foundation's free consulting service in the design of group-randomized trials, at http://sitemaker.umich.edu/group-based/consultation_service; Steve Raudenbush et. al., *Optimal Design Software for Group Randomized Trials*, at http://sitemaker.umich.edu/group-based/optimal_design_software; Peter Z. Schochet, *Statistical Power for Random Assignment Evaluations of Education Programs* (http://www.mathematica-mpr.com/publications/PDFs/statisticalpower.pdf), prepared for the U.S. Education Department's Institute of Education Sciences, June 22, 2005; and Howard S. Bloom, "Randomizing Groups to Evaluate Place-Based Programs," in *Learning More from Social Experiments: Evolving Analytical Approaches*, edited by Howard S. Bloom. New York: Russell Sage Foundation Publications, 2005, pp. 115-172.

[5] Here are illustrative examples of sample sizes from well-conducted randomized controlled trials in various areas of social policy: (i) 4,028 welfare applicants and recipients were randomized in a trial of Portland Oregon's Job Opportunities and Basic Skills Training Program (a welfare-to work program), to evaluate the program's effects on employment and earnings – see http://evidencebasedprograms.org/wordpress/?page_id=140; (ii) between 400 and 800 women were randomized in each of three trials of the Nurse-Family Partnership (a nurse home visitation program for low-income, pregnant women), to evaluate the program's effects on a range of maternal and child outcomes, such as child abuse and neglect, criminal arrests, and welfare dependency – see http://evidencebasedprograms.org/wordpress/?page_id=57; 206 9th graders were randomized in a trial of Check and

Connect (a school dropout prevention program for at-risk students), to evaluate the program's effects on dropping out of school – see http://evidencebasedprograms.org/wordpress/?page_id=92; 56 schools containing nearly 6000 students were randomized in a trial of LifeSkills Training (a substance-abuse prevention program), to evaluate the program's effects on students' use of drugs, alcohol, and tobacco – see http://evidencebasedprograms.org/wordpress/?page_id=128.

[6] The study, after obtaining estimates of the intervention's effect with sample members kept in their original groups, can sometimes use a "no-show" adjustment to estimate the effect on intervention group members who actually participated in the intervention (as opposed to no-shows). A variation on this technique can sometimes be used to adjust for "cross-overs." See Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods*, Sage Publications, Inc., 1999, p. 62 and 210; and Howard S. Bloom, "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, vol. 8, April 1984, pp. 225-246.

[7] Similarly, a study of a crime prevention program that involves close police supervision of program participants should not use arrest rates as a measure of criminal outcomes, because the supervision itself may lead to more arrests for the intervention group.

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Evaluating Social Programs Course: Evaluation Glossary

(Sources: 3ie and The World Bank)

Attribution

The extent to which the observed change in outcome is the result of the intervention, having allowed for all other factors which may also affect the outcome(s) of interest.

Attrition

Either the drop out of subjects from the sample during the intervention, or failure to collect data from a subject in subsequent rounds of a data collection. Either form of attrition can result in biased impact estimates.

Baseline

Pre-intervention, ex-ante. The situation prior to an intervention, against which progress can be assessed or comparisons made. Baseline data are collected before a program or policy is implemented to assess the “before” state.

Bias

The extent to which the estimate of impact differs from the true value as a result of problems in the evaluation or sample design.

Cluster

A cluster is a group of subjects that are similar in one way or another. For example, in a sampling of school children, children who attend the same school would belong to a cluster, because they share the same school facilities and teachers and live in the same neighborhood.

Cluster sample

Sample obtained by drawing a random sample of clusters, after which either all subjects in selected clusters constitute the sample or a number of subjects within each selected cluster is randomly drawn.

Comparison group

A group of individuals whose characteristics are similar to those of the treatment groups (or participants) but who do not receive the intervention. Comparison groups are used to approximate the counterfactual. In a randomized evaluation, where the evaluator can ensure that no confounding factors affect the comparison group, it is called a control group.

Confidence level

The level of certainty that the true value of impact (or any other statistical estimate) will fall within a specified range.

Confounding factors

Other variables or determinants that affect the outcome of interest.

Contamination

When members of the control group are affected by either the intervention (see “spillover effects”) or another intervention that also affects the outcome of interest. Contamination is a common problem as there are multiple development interventions in most communities.

Cost-effectiveness

An analysis of the cost of achieving a one unit change in the outcome. The advantage compared to cost-benefit analysis, is that the (often controversial) valuation of the outcome is avoided. Can be used to compare the relative efficiency of programs to achieve the outcome of interest.

Counterfactual

The counterfactual is an estimate of what the outcome would have been for a program participant in the absence of the program. By definition, the counterfactual cannot be observed. Therefore it must be estimated using comparison groups.

Dependent variable

A variable believed to be predicted by or caused by one or more other variables (independent variables). The term is commonly used in regression analysis.

Difference-in-differences (also known as double difference or D-in-D)

The difference between the change in the outcome in the treatment group compared to the equivalent change in the control group. This method allows us to take into account any differences between the treatment and comparison groups that are constant over time. The two differences are thus before and after and between the treatment and comparison groups.

Evaluation

Evaluations are periodic, objective assessments of a planned, ongoing or completed project, program, or policy. Evaluations are used to answer specific questions often related to design, implementation and/or results.

Ex ante evaluation design

An impact evaluation design prepared before the intervention takes place. Ex ante designs are stronger than ex post evaluation designs because of the possibility of considering random assignment, and the collection of baseline data from both treatment and control groups. Also called prospective evaluation.

Ex post evaluation design

An impact evaluation design prepared once the intervention has started, and possibly been completed. Unless the program was randomly assigned, a quasi-experimental design has to be used.

External validity
The extent to which the causal impact discovered in the impact evaluation can be generalized to another time, place, or group of people. External validity increases when the evaluation sample is representative of the universe of eligible subjects.

Follow-up survey
Also known as “post-intervention” or “ex-post” survey. A survey that is administered after the program has started, once the beneficiaries have benefited from the program for some time. An evaluation can include several follow-up surveys.

Hawthorne effect
The “Hawthorne effect” occurs when the mere fact that you are observing subjects makes them behave differently.

Hypothesis
A specific statement regarding the relationship between two variables. In an impact evaluation the hypothesis typically relates to the expected impact of the intervention on the outcome.

Impact
The effect of the intervention on the outcome for the beneficiary population.

Impact evaluation
An impact evaluation tries to make a causal link between a program or intervention and a set of outcomes. An impact evaluation tries to answer the question of whether a program is responsible for changes in the outcomes of interest. Contrast with “process evaluation”.

Independent variable
A variable believed to cause changes in the dependent variable, usually applied in regression analysis.

Indicator
An indicator is a variable that measures a phenomenon of interest to the evaluator. The phenomenon can be an input, an output, an outcome, or a characteristic.

Inputs
The financial, human, and material resources used for the development intervention.

Intention to treat (ITT) estimate
The average treatment effect calculated across the whole treatment group, regardless of whether they actually participated in the intervention or not. Compare to “treatment on the treated estimate”.

Intra-cluster correlation
Intra-cluster correlation is correlation (or similarity) in outcomes or characteristics between subjects that belong to the same cluster. For example, children that attend the

same school would typically be similar or correlated in terms of their area of residence or socio-economic background.

Logical model
Describes how a program should work, presenting the causal chain from inputs, through activities and outputs, to outcomes. While logical models present a theory about the expected program outcome, they do not demonstrate whether the program caused the observed outcome. A theory-based approach examines the assumptions underlying the links in the logical model.

John Henry effect
The "John Henry effect" happens when comparison subjects work harder to compensate for not being offered a treatment. When one compares treated units to those "harder-working" comparison units, the estimate of the impact of the program will be biased: we will estimate a smaller impact of the program than the true impact we would find if the comparison units did not make the additional effort.

Minimum desired effect
Minimum change in outcomes that would justify the investment that has been made in an intervention, accounting not only for the cost of the program and the type of benefits that it provides, but also on the opportunity cost of not having invested funds in an alternative intervention. The minimum desired effect is an input for power calculations: evaluation samples need to be large enough to detect at least the minimum desired effects with sufficient power.

Null hypothesis
A null hypothesis is a hypothesis that might be falsified on the basis of observed data. The null hypothesis typically proposes a general or default position. In evaluation, the default position is usually that there is no difference between the treatment and control group, or in other words, that the intervention has no impact on outcomes.

Outcome
A variable that measures the impact of the intervention. Can be intermediate or final, depending on what it measures and when.

Output
The products and services that are produced (supplied) directly by an intervention. Outputs may also include changes that result from the intervention which are relevant to the achievement of outcomes.

Power calculation
A calculation of the sample required for the impact evaluation, which depends on the minimum effect size that we want to be able to detect (see “minimum desired effect”) and the required level of confidence.

Pre-post comparison
Also known as a before and after comparison. A pre-post comparison attempts to establish the impact of a program by tracking changes in outcomes for program beneficiaries over time using measures both before and after the program or policy is implemented.

Process evaluation
A process evaluation is an evaluation that tries to establish the level of quality or success of the processes of a program. For example: adequacy of the administrative processes, acceptability of the program benefits, clarity of the information campaign, internal dynamics of implementing organizations, their policy instruments, their service delivery mechanisms, their management practices, and the linkages among these. Contrast with “impact evaluation”.

Quasi-experimental design
Impact evaluation designs that create a control group using statistical procedures. The intention is to ensure that the characteristics of the treatment and control groups are identical in all respects, other than the intervention, as would be the case in an experimental design.

Random assignment
An intervention design in which members of the eligible population are assigned at random to either the treatment group (receive the intervention) or the control group (do not receive the intervention). That is, whether someone is in the treatment or control group is solely a matter of chance, and not a function of any of their characteristics (either observed or unobserved).

Random sample
The best way to avoid a biased or unrepresentative sample is to select a random sample. A random sample is a probability sample where each individual in the population being sampled has an equal chance (probability) of being selected.

Randomized evaluation (RE) (also known as randomized controlled trial, or RCT)
An impact evaluation design in which random assignment is used to allocate the intervention among members of the eligible population. Since there should be no correlation between participant characteristics and the outcome, and differences in outcome between the treatment and control can be fully attributed to the intervention, i.e. there is no selection bias. However, REs may be subject to several types of bias and so need follow strict protocols. Also called “experimental design”.

Regression analysis
A statistical method which determines the association between the dependent variable and one or more independent variables.

Selection bias
A possible bias introduced into a study by the selection of different types of people into treatment and comparison groups. As a result, the outcome differences may potentially be explained as a result of pre-existing differences between the groups, rather than the treatment itself.

Significance level
The significance level is usually denoted by the Greek symbol, α (alpha). Popular levels of significance are 5% (0.05), 1% (0.01) and 0.1% (0.001). If a test of significance gives a p-value lower than the α-level, the null hypothesis is rejected. Such results are informally referred to as 'statistically significant'. The lower the significance level, the stronger the evidence required. Choosing level of significance is an arbitrary task, but for many applications, a level of 5% is chosen, for no better reason than that it is conventional.

Spillover effects
When the intervention has an impact (either positive or negative) on units not in the treatment group. Ignoring spillover effects results in a biased impact estimate. If there are spillover effects then the group of beneficiaries is larger than the group of participants.

Stratified sample
Obtained by dividing the population of interest (sampling frame) into groups (for example, male and female), then by drawing a random sample within each group. A stratified sample is a probabilistic sample: every unit in each group (or strata) has the same probability of being drawn.

Treatment group
The group of people, firms, facilities or other subjects who receive the intervention. Also called participants.

Treatment on the treated (TOT) estimate
The treatment on the treated estimate is the impact (average treatment effect) only on those who actually received the intervention. Compare to intention to treat.

Unobservables
Characteristics which cannot be observed or measured. The presence of unobservables can cause selection bias in quasi-experimental designs.